

# Needles and straw in a haystack: robust empirical Bayes confidence for possibly sparse sequences

BY EDUARD BELITSER AND NURZHAN NURUSHEV

*VU Amsterdam*

April 5, 2016

In the many normal means model we construct an empirical Bayes posterior which we then use for *uncertainty quantification* for the unknown (possibly sparse) parameter by constructing an estimator and a confidence set around it as empirical Bayes credible ball. We allow the model to be misspecified (the normality assumption can be dropped, with some moment conditions instead), leading to the robust empirical Bayes inference. An important step in assessing the uncertainty is the derivation of the fact that the empirical Bayes posterior contracts to the parameter with a local (i.e., depending on the parameter) rate which is the best over certain family of local rates; therefore called *oracle rate*. We introduce the so called *excessive bias restriction* under which we establish the local (oracle) confidence optimality of the empirical Bayes credible ball. Adaptive minimax results (for the estimation and posterior contraction problems) over sparsity classes follow from our local results. An extra (square root of) log factor appears in the radial rate of the confidence ball; it is not known whether this is an artifact or not.

## 1 Introduction

Suppose we observe  $X = X^{(\sigma, n)} = (X_1, \dots, X_n)$ :

$$X_i = \theta_i + \sigma \xi_i, \quad i \in \mathbb{N}_n = \{1, \dots, n\}, \quad (1)$$

where  $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$  is an unknown high-dimensional parameter of interest, the  $\xi_i$ 's are independent errors with  $E\xi_i = 0$  and  $\text{Var}(\xi_i) = 1$ , and the (known) noise intensity  $\sigma > 0$ . The leading case will be Gaussian:  $\xi_i \sim N(0, 1)$ ; but we also allow the model to be *misspecified*. That is, the  $\xi_i$ 's may be not necessarily normal, but satisfying certain moment conditions instead; see subsection 4.1. The general goal is to make inference

---

*MSC 2010 subject classification:* primary 62G15, 62C12.

*Keywords and phrases:* empirical Bayes posterior, optimal confidence set, oracle contraction rate.

about the parameter  $\theta$  based on the data  $X$ : recovery of  $\theta$  and *uncertainty quantification* by constructing an *optimal confidence set*. We will make sense of these notions later.

We measure the size of a confidence set by the smallest radius of a ball containing this set, hence it is enough to consider confidence balls. For the usual norm  $\|\cdot\|$  in  $\mathbb{R}^n$ , a random ball in  $\mathbb{R}^n$  is  $B(\hat{\theta}, \hat{R}) = \{\theta \in \mathbb{R}^n : \|\hat{\theta} - \theta\| \leq \hat{R}\}$ , where the center  $\hat{\theta} = \hat{\theta}(X) : \mathbb{R}^n \mapsto \mathbb{R}^n$  and radius  $\hat{R} = \hat{R}(X) : \mathbb{R}^n \mapsto \mathbb{R}_+ = [0, +\infty]$  are measurable functions of the data  $X$ . Next, we consider mainly the non-asymptotic results, which imply asymptotic assertions if needed. There are following asymptotic regimes: decreasing noise level  $\sigma \rightarrow 0$ , (the leading case) high-dimensional setup  $n \rightarrow \infty$ , or their combination, e.g.,  $\sigma = n^{-1/2}$  and  $n \rightarrow \infty$ .

Useful inference is not possible without some structure on the parameter  $\theta$ . Popular structural assumptions are *smoothness* and *sparsity*. Smoothness assumption is suitable when the structure on the parameter  $\theta$  is imposed in terms of smoothness classes (e.g., a scale of Sobolev ellipsoids). In this paper, we are concerned with *sparse* parameters, i.e., with a large proportion of zero (or any other known value) coordinates. In the literature this structure is usually modeled by the global sparsity classes such as *nearly black vectors* or *weak  $\ell_s$ -balls*; see the definitions in Subsection 3.4.

The best studied problem in the sparsity context is that of estimating  $\theta$  in the many normal means model, a variety of estimation methods and results are available in the literature: Donoho and Johnstone (1994b), Birgé and Massart (2001), Johnstone and Silverman (2004), Abramovich, Benjamini, Donoho and Johnstone (2006), Abramovich, Grinshtein and Pensky (2007), Castillo and van der Vaart (2012), van der Pas, Kleijn and van der Vaart (2014). However, even an optimal estimator does not reveal how far it is from  $\theta$ . It is of great importance to quantify this uncertainty, which can be cast into the problem of constructing confidence sets for the parameter  $\theta$ .

Many inference methods have Bayesian connections. For example, even some seemingly non Bayesian estimators can be obtained as certain quantities (like posterior mode for penalized minimum contrast estimators) of the (empirical Bayes) posterior distributions resulting from imposing some specific priors on the parameter; cf. Johnstone and Silverman (2004) and Abramovich, Grinshtein and Pensky (2007). Although the Bayesian methodology is used or can be related to in constructing many (frequentist) inference procedures, only recently the posterior distributions themselves have been studied for the model (1) in the sparsity context: Castillo and van der Vaart (2012), van der Pas, Kleijn and van der Vaart (2014). The quality of posterior is characterized by the posterior contraction rate around the true parameter. These are usually related to the minimax rates over certain sparsity classes from the corresponding estimation problems.

A common Bayesian way to model sparsity structure is by the so called two-groups priors. Such a prior puts positive mass on vectors  $\theta$  with some exact zero coordinates (zero group) and the remaining coordinates (signal group) are drawn (independently) from a chosen distribution. So the marginal prior for each coordinate is a mixture of a continuous distribution and a point-mass at zero. Castillo and van der Vaart (2012) show that for a suitably chosen two-groups prior, the posterior concentrates around the true  $\theta$  at the minimax rate (as  $n \rightarrow \infty$ ) for two sparsity classes. As pointed out by Castillo and van der Vaart (2012) (and also by Johnstone and Silverman (2004) in the estimation context), densities for non-zero coordinates should not have too light tails, otherwise one

gets sub-optimal concentration properties. The important Gaussian case is for example excluded. This has to do with the so called *over-shrinkage effect* of the normal prior with a fixed mean for nonzero coordinates, which pushes the posterior too much towards the prior mean, missing the true parameter that in general differs from the prior mean. That is why Castillo and van der Vaart (2012) and Johnstone and Silverman (2004) discard normal priors on non-zero coordinates and use heavy tailed priors. A way to construct such a prior is to put a next level heavy-tailed prior, like half-Cauchy, on the variance in the normal prior, resulting in the so called horseshoe prior on  $\theta$  (of course, the normal-normal conjugacy structure is then lost, resulting in one-component prior). This prior was proposed by Carvalho, Polson and Scott (2010) and used by van der Pas, Kleijn and van der Vaart (2014). In the present paper we show that normal priors are still usable and even lead to strong local results if combined with empirical Bayes approach.

The main aim in this paper is to construct, by using the empirical Bayes approach, a confidence ball  $B(\hat{\theta}, C\hat{r})$  such that for any  $\alpha_1, \alpha_2 \in (0, 1]$  and some functional  $r(\theta) = r_{\sigma,n}(\theta)$ ,  $r : \mathbb{R}^n \rightarrow \mathbb{R}_+$ , there exists  $C, c > 0$  such that

$$\sup_{\theta \in \Theta_0} P_\theta(\theta \notin B(\hat{\theta}, C\hat{r})) \leq \alpha_1, \quad \sup_{\theta \in \Theta_1} P_\theta(\hat{r} \geq cr(\theta)) \leq \alpha_2, \quad (2)$$

for some  $\Theta_0, \Theta_1 \subseteq \mathbb{R}^n$ . The functional  $r(\theta)$ , called the *radial rate*, is a benchmark for the effective radius of the confidence ball  $B(\hat{\theta}, C\hat{r})$ . The first expression in (2) is called *coverage relation* and the second *size relation*. It is desirable to find the smallest  $r(\theta)$  and the biggest  $\Theta_0, \Theta_1$ , for which (2) holds. These are contrary requirements, so we have to trade them off against each other. There are different ways of doing this, leading to different optimality frameworks. For example, if we insist on  $\Theta_0 = \Theta_1 = \mathbb{R}^n$ , then the results by Li (1989) and by Cai and Low (2004) say basically that the radial rate  $r$  cannot be of a faster order than  $\sigma n^{1/4}$  for every  $\theta$  and is at least of order  $\sigma n^{1/2}$  for some  $\theta$ . A more refined version of this situation was investigated by Baraud (2004):  $\Theta_1 = S$  for some linear subspace  $S \subseteq \mathbb{R}^n$ , leading to a lower bound for the radial rate in terms of the subspace  $S$ .

Suppose we have a smoothness structure  $\theta \in \Theta_\beta$  with unknown “smoothness”  $\beta \in \mathcal{B}$  (e.g.,  $\mathcal{B} = [\beta_{min}, \beta_{max}]$ ), and  $r(\Theta_\beta)$  is the minimax estimation rate over  $\Theta_\beta$ . Then the minimax adaptive version of (2) would be obtained by taking  $\Theta_0 = \Theta_1 = \Theta_\beta$  and the radial rate  $r(\theta) = r(\Theta_\beta)$  for all  $\theta \in \Theta_\beta$ . However, it turns out that the uncertainty quantification problem (2) is intrinsically more difficult than the estimation and posterior contraction problems. Unlike for the estimation and posterior contraction problems, the minimax adaptive version of (2) is impossible to hold. The description of this phenomenon can be found in Robins and van der Vaart (2006), Bull and Nickl (2013), Belitser (2014) and Szabó, van der Vaart and van Zanten (2015) for some typical smoothness structures, e.g., Sobolev classes. Roughly speaking, the coverage relation in (2) will not hold for all  $\Theta_0 = \Theta_\beta$ , but only for  $\Theta_0 = \Theta_\beta \setminus \Theta'$ , with some set of “deceptive parameters”  $\Theta'$  removed from  $\Theta_\beta$ . Szabó et al. (2015) call such parameters “inconvenient truths” and give an implicit construction of a  $\theta' \in \Theta'$ . Examples of non-deceptive parameters are the set of *self-similar* parameters  $\Theta_0 = \Theta_{ss}$  introduced by Picard and Tribouley (2000) and studied by Bull (2012), Bull and Nickl (2013), Nickl and Szabó (2014), Szabó, van der

Vaart and van Zanten (2015), and the set of *polished tail parameters*  $\Theta_0 = \Theta_{pt}$  considered by Szabó et al. (2015). In all the above mentioned papers global minimax radial rates (i.e.,  $r(\theta) = r(\Theta_\beta)$  for all  $\theta \in \Theta_\beta$ ) for specific smoothness structures were studied. A local approach, delivering also the adaptive minimax results for many smoothness structures simultaneously, was considered by Babenko and Belitser (2010) for posterior contraction rates and by Belitser (2014) for constructing optimal confidence balls. In Belitser (2014), yet a more general (than  $\Theta_{ss}$  and  $\Theta_{pt}$ ) set of non-deceptive parameters was introduced,  $\Theta_0 = \Theta_{ebr}$ , parameters satisfying the so called *excessive bias restriction*.

To the best of our knowledge, there are no adaptive minimax results on uncertainty quantification (2) for sparsity structures, not even for specific sparsity classes. This paper attempts to fill this gap and even to extend to the misspecified models (we allow the model to be not necessarily normal). Moreover, we pursue the novel local approach, namely, the radial rate  $r(\theta)$  in (2) is allowed to be a function of  $\theta$ , which, in a way, measures the amount of sparsity for each  $\theta \in \mathbb{R}^n$ : the smaller  $r(\theta)$ , the more sparse  $\theta$ . The local approach is more powerful than the global one. The point is that then we do not need to impose any specific sparsity structure, because the proposed local approach automatically exploits the “effective” sparsity of each underlying  $\theta$ . If  $\theta$  happens to lie in a sparsity class, adaptive (global) minimax results (in fact, for the three problems: estimation, posterior contraction rate and confidence sets) over this sparsity class follow from the local results. In particular, our local results imply the same type of certain (global) minimax estimation results over sparsity classes as in Johnstone and Silverman (2004), and the same type of global minimax (over sparsity classes) results on contraction posterior rates as in Castillo and van der Vaart (2012).

The paper is organized as follows. In Section 2 we introduce the notations, the prior and describe the empirical Bayes procedure in detail. In Section 3, we obtain the upper bound result for the proposed empirical Bayes posterior in terms of the local radial rate  $r(\theta_0)$  uniformly over  $\mathbb{R}^n$ . This result, besides being an ingredient for the uncertainty quantification problem (2), is of importance on its own as it actually establishes the contraction of the empirical Bayes posterior with the local (oracle) rate at least  $r(\theta_0)$ . Global minimax (over sparsity classes) results on contraction posterior rates follow. As another consequence, we also obtain the oracle estimation result by constructing an estimator, the empirical Bayes posterior mean, which converges to  $\theta_0$  with the local (oracle) rate  $r(\theta_0)$ . This local estimation result in turn implies (global) minimax estimation results over sparsity classes.

Next, we propose the general construction of confidence ball by using empirical Bayes posterior, as empirical Bayes credible ball. Then we derive the lower bound result, which, roughly speaking, means that the empirical Bayes posterior contracts around the empirical Bayes posterior mean with a rate at most  $r(\theta_0)$ . The lower bound result holds uniformly only over some set of parameters satisfying the so called *excessive bias restriction* (we use the same name as a similar condition in Belitser (2014) for the smoothness structures),  $\Theta_0 = \Theta_{eb} \subseteq \mathbb{R}^n$ , which forms an actual restriction. It is in accordance with the above mentioned fact that it is not possible to construct optimal (fully) adaptive confidence set in the minimax sense and in terms of local radial rate. Combining the upper and lower bound results, we finally derive the optimality (2) of our credible ball with  $\Theta_0 = \Theta_{eb}$ ,

$\Theta_1 = \mathbb{R}^n$  and the local radial rate  $(\log n)r(\theta_0)$ . We have to mention that an extra (square root of)  $\log n$  factor appears in the radial rate of the resulting confidence ball; it is not known whether this is an artifact or not.

In Section 4, we discuss some extensions and variations of our results and present concluding remarks. In particular, we elaborate on one important extension: namely, we allow the model to be misspecified (the normality assumption can be dropped, with some moment conditions instead), leading to the robust empirical Bayes inference.

The theoretical results are illustrated in Section 5 by a small simulation study. All proofs of lemmas and theorems are gathered in Sections 6 and 7.

## 2 Preliminaries

First we introduce some notations, then multivariate normal prior. Next, by applying the empirical Bayes approach, leading to the empirical Bayes posterior which we will use in the construction of the estimator and the confidence ball as credible ball with respect to this resulting empirical Bayes posterior.

### 2.1 Notations

Denote the probability measure of  $X$  from the model (1) by  $P_\theta = P_\theta^{(\sigma, n)}$ . For the notational simplicity we often skip the dependence on  $\sigma$  and  $n$  of this quantity and many others. Denote by  $1\{s \in S\} = 1_S(s)$  the indicator function of the set  $S$ , by  $|S|$  the cardinality of the set  $S$ , the difference of sets  $S \setminus S_0 = \{s \in S : s \notin S_0\}$ ,  $\mathbb{N}_k = \{1, \dots, k\}$  for  $k \in \mathbb{N} = \{1, 2, \dots\}$ . For  $I \subseteq \mathbb{N}_n$  define  $I^c = \mathbb{N}_n \setminus I$ . Throughout  $Z \sim N(0, 1)$  will denote a generic standard normal random variable, with distribution function  $\Phi(z) = P(Z \leq z)$  and density  $\phi(z) = \Phi'(z)$ . Let  $\phi(x, \mu, \sigma^2)$  be the density of  $\mu + \sigma Z \sim N(\mu, \sigma^2)$  at point  $x$ . By convention,  $N(\mu, 0) = \delta_\mu$  denotes a Dirac measure at point  $\mu$ .

Consider the family  $\mathcal{M} = \mathcal{M}_n$  of all subsets of  $\mathbb{N}_n$  except for the empty set, i.e.,  $\mathcal{M} = 2^{\mathbb{N}_n} \setminus \emptyset = \{I : I \subseteq \mathbb{N}_n, I \neq \emptyset\}$  and  $|\mathcal{M}| = 2^n - 1$ .

### 2.2 Multivariate normal prior

As we mentioned in the introduction, we deal with the classical high-dimensional normal model  $X = (X_i, i \in \mathbb{N}_n) \sim P_\theta = \bigotimes_{i=1}^n N(\theta_i, \sigma^2)$ ,  $\theta = (\theta_i, i \in \mathbb{N}_n) \in \mathbb{R}^n$ . Suppose there are two distinct groups of coordinates of  $\theta$ : for some  $I \in \mathcal{M}_n$ ,  $\theta_I = (\theta_i, i \in I)$  and  $\theta_{I^c} = (\theta_i, i \in I^c)$ , so that  $\theta = (\theta_I, \theta_{I^c})$ . In this paper we study the sparsity case: the two groups in vector  $\theta$  are (almost) zeros  $\theta_{I^c} = (\theta_i, i \notin I)$  and non-zeros  $\theta_I = (\theta_i, i \in I)$ . Then it is reasonable to impose a prior on  $\theta$  given  $I$  as follows:

$$\pi_I = \bigotimes_{i=1}^n N(\mu_i(I), \tau_i^2(I)), \quad (3)$$

$$\mu_i(I) = \mu_{1,i} 1\{i \in I\}, \quad \tau_i^2(I) = \sigma^2 K_n(I) 1\{i \in I\}, \quad K_n(I) = \frac{ne}{|I|} - 1. \quad (4)$$

This rather specific choice of  $K_n(I)$  is made for the sake of concise expressions in later calculations, many other choices are actually possible.

Recall the elementary fact: if  $X|\theta \sim N(\theta, \sigma^2)$  and  $\theta \sim N(\mu, \tau^2)$ , then

$$\theta|X \sim N(\mu_X, \tau_X^2), \quad \mu_X = \frac{\tau^2 X + \sigma^2 \mu}{\tau^2 + \sigma^2}, \quad \tau_X^2 = \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2}.$$

In view of this fact and the product structure, the corresponding posterior distribution for  $\theta$  is readily obtained:

$$\pi_I(\theta|X) = \bigotimes_{i=1}^n N\left(\frac{\tau_i^2(I)X_i + \sigma^2 \mu_i(I)}{\tau_i^2(I) + \sigma^2}, \frac{\tau_i^2(I)\sigma^2}{\tau_i^2(I) + \sigma^2}\right). \quad (5)$$

Next, introduce the prior  $\lambda$  on  $I \in \mathcal{M}_n$  (see Subsection 4.3 below):

$$\lambda(\mathcal{I} = I) = \lambda_I = c_{\kappa,n} \exp\left\{-\kappa|I| \log\left(\frac{en}{|I|}\right)\right\} = c_{\kappa,n} \left(\frac{en}{|I|}\right)^{-\kappa|I|}, \quad (6)$$

with  $\kappa > 1$ . Since  $\exp\{l \log(\frac{n}{l})\} \leq \binom{n}{l} \leq \exp\{l \log(\frac{en}{l})\}$ , we evaluate

$$\begin{aligned} 1 &= \sum_{I \in \mathcal{M}_n} \lambda_I = c_{\kappa,n} \sum_{l=1}^n \binom{n}{l} \exp\left\{-\kappa l \log(en/l)\right\} \\ &\leq c_{\kappa,n} \sum_{l=1}^n \exp\{l \log(\frac{en}{l})\} \exp\left\{-\kappa l \log(\frac{en}{l})\right\} \leq c_{\kappa,n} \sum_{l=1}^n e^{-(\kappa-1)l}, \end{aligned}$$

so that  $c_{\kappa,n} \geq e^{\kappa-1} - 1 > 0$  for all  $n \in \mathbb{N}$ . Combining (3) and (6) gives the mixture prior on  $\theta$ :  $\pi = \sum_{I \in \mathcal{M}_n} \lambda_I \pi_I$ . This leads to the marginal distribution of  $X$

$$P_X = \sum_{I \in \mathcal{M}_n} \lambda_I P_{X,I}, \quad P_{X,I} = \bigotimes_{i=1}^n N(\mu_i(I), \sigma^2 + \tau_i^2(I)),$$

and the posterior of  $\theta$

$$\pi(\theta|X) = \sum_{I \in \mathcal{M}_n} \pi(\theta, \mathcal{I} = I|X) = \sum_{I \in \mathcal{M}_n} \pi(\theta|X, \mathcal{I} = I) \pi(\mathcal{I} = I|X), \quad (7)$$

where  $\pi(\theta|X, \mathcal{I} = I) = \pi_I(\theta|X)$  is defined by (5) and the posterior for  $I$  is

$$\pi(\mathcal{I} = I|X) = \frac{\lambda_I P_{X,I}}{P_X} = \frac{\lambda_I \prod_{i=1}^n \phi(X_i, \mu_i(I), \sigma^2 + \tau_i^2(I))}{\sum_{J \in \mathcal{M}_n} \lambda_J \prod_{i=1}^n \phi(X_i, \mu_i(J), \sigma^2 + \tau_i^2(J))}. \quad (8)$$

### 2.3 Empirical Bayes posterior

The parameters  $\mu_{1,i}$  are yet to be chosen in the prior. We choose  $\mu_{1,i}$  by using empirical Bayes approach. The marginal likelihood  $P_X$  is readily maximized with respect to  $\mu_{1,i}$ :

$\hat{\mu}_{1,i} = X_i$ , which we then substitute instead of  $\mu_{1,i}$  in the expression (7) for  $\pi(\theta|X)$ , obtaining the empirical Bayes posterior

$$\hat{\pi}(\theta|X) = \hat{\pi}_\kappa(\theta|X) = \sum_{I \in \mathcal{M}_n} \hat{\pi}(\theta|X, \mathcal{I} = I) \hat{\pi}(\mathcal{I} = I|X), \quad (9)$$

where the empirical Bayes conditional posterior (recall that  $N(0,0) = \delta_0$ )

$$\begin{aligned} \hat{\pi}(\theta|X, \mathcal{I} = I) &= \hat{\pi}_I(\theta|X) = \left[ \bigotimes_{i \in I} N(X_i, \frac{K_n(I)\sigma^2}{K_n(I)+1}) \right] \bigotimes_{i \in I^c} N(0,0) \\ &= \bigotimes_{i=1}^n N(X_i 1\{i \in I\}, \frac{K_n(I)\sigma^2 1\{i \in I\}}{K_n(I)+1}) \end{aligned} \quad (10)$$

is obtained from (4) and (5) with  $\mu_{1,i} = X_i$ , and

$$\hat{\pi}(\mathcal{I} = I|X) = \hat{\pi}_I = \frac{\lambda_I \prod_{i=1}^n \phi(X_i, X_i 1\{i \in I\}, \sigma^2 + \tau_i^2(I))}{\sum_{J \in \mathcal{M}_n} \lambda_J \prod_{i=1}^n \phi(X_i, X_i 1\{i \in J\}, \sigma^2 + \tau_i^2(J))} \quad (11)$$

is the empirical Bayes posterior for  $I \in \mathcal{M}_n$ , obtained from (4) and (8) with  $\mu_{1,i} = X_i$ .

Denoting  $X(I) = (X_i 1\{i \in I\}, i \in \mathbb{N}_n)$ , introduce an estimator

$$\hat{\theta} = \hat{\theta}(I) = E_{\hat{\pi}}(\theta|X) = \sum_{I \in \mathcal{M}_n} X(I) \hat{\pi}(\mathcal{I} = I|X), \quad (12)$$

which is nothing else but the *empirical Bayes posterior mean*.

Consider an alternative empirical Bayes posterior. First derive an empirical Bayes variable selector  $\check{I}$  by maximizing  $\hat{\pi}(\mathcal{I} = I|X)$  over  $I \in \mathcal{M}_n$ :

$$\begin{aligned} \check{I} = \operatorname{argmax}_{I \in \mathcal{M}_n} \hat{\pi}(\mathcal{I} = I|X) &= \operatorname{argmax}_{I \in \mathcal{M}_n} \left\{ - \sum_{i \in I^c} \frac{X_i^2}{2\sigma^2} - \frac{|I|}{2} \log(K_n(I) + 1) + \log \lambda_I \right\} \\ &= \operatorname{argmin}_{I \in \mathcal{M}_n} \left\{ - \sum_{i \in I} X_i^2 + (2\kappa + 1)|I|\sigma^2 \log\left(\frac{en}{|I|}\right) \right\}. \end{aligned} \quad (13)$$

Plugging in this into  $\hat{\pi}_I(\theta|X)$  defined by (10) gives the corresponding empirical (now “twice empirical”: with respect to  $\mu_{1,i}$  and with respect to  $I$ ) Bayes posterior, yielding also the empirical Bayes mean estimator for  $\theta$ :

$$\check{\pi}(\theta|X) = \hat{\pi}_{\check{I}}(\theta|X), \quad \check{\theta} = \check{\theta}(\check{I}) = E_{\check{\pi}}(\theta|X) = X(\check{I}). \quad (14)$$

### 3 Main results

In this section we give the main results of the paper.

### 3.1 Empirical Bayes posterior contraction with the oracle rate

First we introduce the local contraction rate for the empirical Bayes posterior  $\hat{\pi}(\cdot|X)$ , which is a random mixture over  $\hat{\pi}_I(\cdot|X)$ ,  $I \in \mathcal{M}_n$ . From the  $P_{\theta_0}$ -perspective, each  $\hat{\pi}_I(\cdot|X)$  contracts to the true parameter  $\theta_0$  with the local rate

$$R^2(I, \theta_0) = R^2(I, \theta_0, \sigma) = \sum_{i \in I^c} \theta_{0,i}^2 + \sigma^2 |I|, \quad I \in \mathcal{M}_n. \quad (15)$$

Indeed,  $X(I) = (X_i 1\{i \in I\}, i \in \mathbb{N}_n)$ , (10) and the Markov inequality yield

$$\begin{aligned} \mathbb{E}_{\theta_0} \hat{\pi}_I(\|\theta - \theta_0\|^2 \geq M^2 R^2(I, \theta_0) | X) &\leq \frac{\mathbb{E}_{\theta_0} (\|X(I) - \theta_0\|^2 + \frac{K_n(I)\sigma^2|I|}{K_n(I)+1})}{M^2 R^2(I, \theta_0)} \\ &= \frac{(1 + \frac{K_n(I)}{K_n(I)+1})\sigma^2|I| + \sum_{i \in I^c} \theta_{0,i}^2}{M^2 R^2(I, \theta_0)} \leq \frac{2}{M^2}. \end{aligned}$$

For each  $\theta_0 \in \mathbb{R}^n$ , there exists the best choice  $I_o = I_o(\theta_0) = I_o(\theta_0, \sigma)$  of the set  $I \in \mathcal{M}_n$  corresponding to the fastest local rate over the family of local rates  $R^2(\mathcal{M}_n) = \{R^2(I, \theta_0), I \in \mathcal{M}_n\}$ : with  $R^2(I, \theta_0)$  defined by (15),

$$R^2(\theta_0) = \min_{I \in \mathcal{M}_n} R^2(I, \theta_0) = \sum_{i \in I_o^c} \theta_{0,i}^2 + \sigma^2 |I_o| = \sigma^2 + \sum_{i=2}^n \min\{\theta_{0,[i]}^2, \sigma^2\},$$

where  $\theta_{0,[1]}^2 \geq \theta_{0,[2]}^2 \geq \dots \geq \theta_{0,[n]}^2$  are the ordered values of  $(\theta_{0,1}^2, \dots, \theta_{0,n}^2)$ .

Ideally, we would like the quantity  $R^2(\theta_0)$  to be the benchmark for the contraction rate of the empirical Bayes posterior  $\hat{\pi}(\cdot|X)$  defined by (9). However, the rate  $R^2(\theta_0) = \sigma^2 + \sum_{i=2}^n \min\{\theta_{0,[i]}^2, \sigma^2\} \leq \sigma^2 + \sum_{i=1}^n \min\{\theta_{0,i}^2, \sigma^2\}$  is unachievable uniformly in  $\theta_0 \in \mathbb{R}^n$ , which is also confirmed by following estimation result of Donoho and Johnstone (1994a):

$$\liminf_{n \rightarrow \infty} \frac{1}{\log n} \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^n} \left[ \frac{\mathbb{E}_{\theta_0} \|\theta - \hat{\theta}\|^2}{\sigma^2 + \sum_{i=1}^n \min\{\theta_{0,i}^2, \sigma^2\}} \right] \geq 2,$$

where the infimum is taken over all estimators, measurable functions of  $X$ . This shows that a reasonable benchmark for the contraction rate must contain a logarithmic factor, as was also shown by Birgé and Massart (2001) for the estimation problem. For a parameter  $\tau \geq 1$ , introduce the family of the so called  $\tau$ -local rates with a logarithmic factor:

$$r_\tau^2(I, \theta_0) = r_\tau^2(I, \theta_0, \sigma, n) = \sum_{i \in I^c} \theta_{0,i}^2 + \tau \sigma^2 |I| \log \left( \frac{en}{|I|} \right), \quad I \in \mathcal{M}_n.$$

Notice that the logarithmic factor is only present in the variance part of the local rate and it is  $\log(\frac{en}{|I|})$ , not  $\log n$ ; cf. Birgé and Massart (2001). There exists the best choice  $I_o^\tau = I_o^\tau(\theta_0) = I_o^\tau(\theta_0, \sigma, n) \in \mathcal{M}_n$  such that

$$r_\tau^2(\theta_0) = r_\tau^2(I_o^\tau, \theta_0) = \min_{I \in \mathcal{M}_n} r_\tau^2(I, \theta_0) = \sum_{i \notin I_o^\tau} \theta_{0,i}^2 + \tau \sigma^2 |I_o^\tau| \log \left( \frac{en}{|I_o^\tau|} \right). \quad (16)$$



We call  $I_o^\tau$  the  $\tau$ -oracle and the quantity  $r_\tau^2(\theta_0)$   $\tau$ -oracle rate. In what follows, for  $\tau = 1$  we denote:  $r^2(\theta_0) = r_1^2(\theta_0)$  called the *oracle rate*, and  $I_o = I_o^1$  called the *oracle*. Since the empty set is not allowed to be the oracle set,  $|I_o| \geq 1$  and  $r^2(\theta_0) \geq \sigma^2$ . Notice that  $I_o^{\tau_1} \subseteq I_o^{\tau_2}$  if  $\tau_1 \geq \tau_2$ , because the function  $f(x) = x \log(ne/x)$  is increasing on  $(0, n]$ . Besides,  $r^2(\theta_0) \leq r_\tau^2(\theta_0) \leq \tau r^2(\theta_0)$  since  $\tau \geq 1$ .

The following theorem establishes a rather powerful property of the empirical Bayes posterior  $\hat{\pi}(\theta|X)$ : from the  $P_{\theta_0}$ -perspective, it contracts to  $\theta_0$  with the oracle rate  $r(\theta_0)$ .

**Theorem 1.** *Let the empirical Bayes posterior  $\hat{\pi}_\kappa(\theta|X)$  and the oracle rate  $r(\theta_0)$  be defined by (9) and (16), respectively, with  $\kappa > 3.27$ . Then there exists a constant  $C_{or} = C_{or}(\kappa) > 0$  such that, for any  $\theta_0 \in \mathbb{R}^n$ ,  $M > 0$ ,*

$$\mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \theta_0\| \geq Mr(\theta_0)|X) \leq \frac{C_{or}}{M^2}.$$

The next theorem claims that the estimator defined by (12) possesses the oracle property for the estimation problem.

**Theorem 2.** *Let the conditions of Theorem 1 be fulfilled and  $\hat{\theta}$  be defined by (12). Then there exists  $C_{est} = C_{est}(\kappa) > 0$  such that for all  $\theta_0 \in \mathbb{R}^n$*

$$\mathbb{E}_{\theta_0} \|\hat{\theta} - \theta_0\|^2 \leq C_{est} r^2(\theta_0).$$

### 3.2 Lower bound for the contraction rate

In the previous subsection we established the upper bound, Theorem 1, for the contraction rate of the empirical Bayes posterior  $\hat{\pi}(\theta|X)$  defined by (9). This is a necessary ingredient for solving the uncertainty quantification problem (2), but not the only one. We also need a lower bound on the contraction rate of the empirical Bayes posterior  $\hat{\pi}(\theta|X)$ , not around the true parameter, but around the estimator  $\hat{\theta}$  which we are going to use as the center of the resulting confidence ball. As is shown by Belitser (2014), such a lower bound is in some sense a minimal condition for any data dependent measure (in particular, empirical Bayes posteriors) used for constructing confidence sets (as credible balls) that satisfy (2).

The below theorem gives a lower bound on the contraction rate of the empirical Bayes posterior  $\hat{\pi}(\theta|X)$  around the estimator  $\hat{\theta}$  defined by (12), from the  $P_{\theta_0}$ -perspective.

**Theorem 3.** *Let the empirical Bayes posterior  $\hat{\pi}_\kappa(\theta|X)$  be given by (9). Then for any  $\tau > \frac{4\kappa(e+1)+2e}{e-2}$  there exists a constant  $\bar{C}_1 = \bar{C}_1(\kappa, \tau) > 0$  such that, for any  $\theta_0 \in \mathbb{R}^n$ , any estimator  $\tilde{\theta} = \tilde{\theta}(X)$  and any  $\delta \in (0, \bar{C}_2]$ ,*

$$\mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \tilde{\theta}\| \leq \delta \sigma |I_o^\tau|^{1/2} |X) \leq \bar{C}_1 \delta [\log(\delta^{-1})]^{1/2},$$

where the  $\tau$ -oracle  $I_o^\tau = I_o^\tau(\theta_0)$  is defined by (16) and  $\bar{C}_2 = \frac{(e-1)^{1/2}}{e^{3/2}} < 1$ .

Informally, the theorem says that, from  $P_{\theta_0}$ -perspective there is no “leakage” of posterior mass through  $\tilde{\theta}$  with a faster rate than the quantity  $\sigma |I_o^\tau|^{1/2}$ , which is the variance-related term of the  $\tau$ -oracle rate  $r_\tau(\theta_0)$ .

Notice that the above lower bound holds uniformly over  $\theta_0 \in \mathbb{R}^n$ , but only in terms of the quantity  $\sigma|I_o^\tau|^{1/2}$ . To tackle the uncertainty quantification problem (2), we actually need such a lower bound in terms of the whole oracle rate  $r(\theta_0)$  (or the  $\tau$ -oracle rate  $r_\tau(\theta_0)$  as  $r^2(\theta_0) \leq r_\tau^2(\theta_0)$  for any  $\tau \geq 1$ ) instead of just the term  $\sigma|I_o^\tau|^{1/2}$ . By introducing the function

$$t_\tau(\theta_0) = \frac{\sum_{i \notin I_o^\tau} \theta_{0,i}^2}{\sigma^2 |I_o^\tau| \log\left(\frac{en}{|I_o^\tau|}\right)}, \quad \theta_0 \in \mathbb{R}^n, \quad (17)$$

we obtain the following corollary of Theorem 3.

**Corollary 1.** *Let the conditions of Theorem 3 be fulfilled. Then for any  $\theta_0 \in \mathbb{R}^n$ , any estimator  $\tilde{\theta} = \tilde{\theta}(X)$  and any  $\delta \in \left(0, \frac{\bar{C}_2}{[(t_\tau(\theta_0) + \tau) \log(en/|I_o^\tau|)]^{1/2}}\right]$ ,*

$$\mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \tilde{\theta}\| \leq \delta r_\tau(\theta_0) | X) \leq \bar{C}_1 [(t_\tau(\theta_0) + \tau) \log(en/|I_o^\tau|)]^{1/2} \delta [\log(\delta^{-1})]^{1/2},$$

where the  $\tau$ -oracle rate  $r_\tau(\theta_0)$  and the  $\tau$ -oracle  $I_o^\tau$  are defined by (16), function  $t_\tau(\theta_0)$  is defined by (17), and  $\bar{C}_1$  and  $\bar{C}_2$  are defined in Theorem 3.

Although the left hand side of the resulting lower bound is now in terms of the  $\tau$ -oracle rate  $r_\tau(\theta_0)$ , the right hand side is not uniform in  $\theta_0 \in \mathbb{R}^n$  anymore, as the function  $t_\tau(\theta_0)$  is in general not bounded uniformly  $\mathbb{R}^n$ . This is in accordance with the fact mentioned in the introduction that it is impossible to construct optimal (fully) adaptive confidence set in the minimax sense with a prescribed high coverage probability. The same problem should occur (and it does) also for our local approach, because otherwise we would have solved the minimax adaptive global problem. Roughly speaking, there exist “deceptive” parameters  $\theta_0 \in \mathbb{R}^n \setminus \Theta_0$  for which the coverage property in (2) does not hold for arbitrarily small  $\alpha_1$ . A solution to this problem is to remove deceptive parameters from  $\mathbb{R}^n$  and consider the remaining set  $\Theta_0$  of non-deceptive ones. Such sets have been introduced in the literature only for certain smoothness structures: examples mentioned in Introduction are *self-similar parameters*, a more general class of *polished tail sequences*, and yet a more general class defined by the *excessive bias restriction* (EBR).

Now we introduce the condition defining the set of non-deceptive parameters for the sparsity structure studied in this paper. As this condition is very much in spirit of EBR introduced by Belitser (2014) (for a certain smoothness structure), we use the same name for it. Define the *excessive bias restriction* (abbreviated as EBR): for  $t > 0$  and  $\tau \geq 1$ ,

$$\Theta_{eb}(t) = \Theta_{eb}(t, \tau) = \{\theta \in \mathbb{R}^n : t_\tau(\theta) \leq t\},$$

where  $I_o^\tau = I_o^\tau(\theta_0)$  and  $t_\tau(\theta)$  are defined by (16) and (17), respectively. The parameters from  $\Theta_{eb}(t)$  are “typical” in a sense that the  $\tau$ -oracle bias is not of a bigger order than the  $\tau$ -oracle variance so that the lower bound from Corollary 1 extends to the whole  $\tau$ -oracle rate  $r_\tau(\theta_0)$  and becomes uniform in  $\theta_0 \in \Theta_{eb}(t)$ .

### 3.3 Confidence ball under excessive bias restriction

Here we present a general construction of a confidence ball by using the estimator  $\hat{\theta} = \hat{\theta}(X)$  defined by (12) and the empirical Bayes posterior  $\hat{\pi}(\theta|X)$  defined by (9). To avoid complicated notations, in this subsection let us simply fix some values  $\kappa > 3.27$ ,  $\tau > \frac{4\kappa(e+1)+2e}{e-2}$ , (e.g.,  $\kappa = 4$ ,  $\tau = 177$ ), so that the conditions of Theorem 1 and Corollary 1 are fulfilled.

For a  $\gamma \in (0, 1)$  (let it be also fixed, say,  $\gamma = 1/2$ ), define the DD-radius

$$\hat{r} = \hat{r}(\gamma, X, \hat{\theta}) = \inf\{r : \hat{\pi}(\|\theta - \hat{\theta}\| \leq r|X) \geq 1 - \gamma\}, \quad (18)$$

and then, for an  $M > 0$ , construct the confidence set

$$C(\hat{\theta}, M\hat{r}) = \left\{ \theta : \left[ \log \left( \frac{en}{|I_o^\tau(\theta)|} \right) \right]^{-1/2} \|\theta - \hat{\theta}\| \leq M\hat{r} \right\}, \quad (19)$$

where the  $\tau$ -oracle  $I_o^\tau(\theta)$  is defined by (16). This set is not a ball, but  $C(\hat{\theta}, M\hat{r}) \subseteq B(\hat{\theta}, (1 + \log n)^{1/2} M\hat{r})$ , i.e., it is a subset of the ball

$$B(\hat{\theta}, (1 + \log n)^{1/2} M\hat{r}) = \{\theta \in \Theta : \|\theta - \hat{\theta}\| \leq (1 + \log n)^{1/2} M\hat{r}\}.$$

The next result, which is the main result in this paper, establishes the coverage and size properties (2) for the defined credible set with  $\Theta_1 = \mathbb{R}^n$ ,  $\Theta_0 = \Theta_{eb}(t)$ , and the “effective” local rate  $\left[ \log \left( \frac{en}{|I_o^\tau(\theta_0)|} \right) \right]^{1/2} r(\theta_0)$ .

**Theorem 4.** *Let the estimator  $\hat{\theta}$  and the confidence set  $C(\hat{\theta}, M\hat{r})$  be defined by (12) and (19), respectively. Then for any  $t > 0$  and any  $\alpha_1, \alpha_2 \in (0, 1)$  there exist  $M_0 = M_0(\alpha_1, t)$  and  $C_0 = C_0(\alpha_2)$  such that, for all  $M \geq M_0$  and  $C \geq C_0$ , the following relations hold*

$$\sup_{\theta_0 \in \Theta_{eb}(t)} \mathbb{P}_{\theta_0}(\theta_0 \notin C(\hat{\theta}, M\hat{r})) \leq \alpha_1, \quad \sup_{\theta_0 \in \mathbb{R}^n} \mathbb{P}_{\theta_0}(\hat{r}^2 \geq Cr^2(\theta_0)) \leq \alpha_2,$$

where the oracle rate  $r^2(\theta_0)$  is defined by (16).

Clearly, Theorem 4 holds with the ball  $B(\hat{\theta}, (1 + \log n)^{1/2} M\hat{r})$  instead of  $C(\hat{\theta}, M\hat{r})$ , the “effective” local rate for this ball is  $(1 + \log n)^{1/2} r(\theta_0)$ .

### 3.4 Implications: the minimax results over sparsity classes

In this subsection we elucidate the potential strength of the oracle approach. In particular, we demonstrate how the global adaptive minimax results over certain scales can be derived from the local results. The point is that if we want to establish global adaptive minimax results over certain scale, say,  $\{\Theta_\beta, \beta \in \mathcal{B}\}$ , with corresponding minimax rates  $\{r(\Theta_\beta), \beta \in \mathcal{B}\}$ , the only thing we need to show is

$$\sup_{\theta_0 \in \Theta_\beta} r^2(\theta_0) \leq cr^2(\Theta_\beta), \quad \forall \beta \in \mathcal{B}.$$

If the above property holds, we say the oracle rate  $r(\theta)$  *covers* the scale  $\{\Theta_\beta, \beta \in \mathcal{B}\}$ . In this case, the local result for the oracle rate  $r(\theta_0)$  will immediately imply the corresponding global adaptive minimax results over the covered scale. Moreover, the adaptive minimax results follow simultaneously for all scales that are covered by the oracle rate  $r(\theta_0)$ .

Next we consider two sparsity scales  $\{\Theta_\beta, \beta \in \mathcal{B}\}$  for which the global minimax results follow from our local results. For simplicity we assume that  $\sigma^2 = 1$ . Let us again (as in the previous subsection) fix some values  $\kappa$  and  $\tau$  such that the conditions of Theorems 1, 2 and 4 are fulfilled.

**Nearly black vectors** For  $p_n \in \mathbb{N}_n$  such that  $p_n = o(n)$  as  $n \rightarrow \infty$ ,

$$\ell_0[p_n] = \{\theta \in \mathbb{R}^n : \#(1 \leq i \leq n : \theta_i \neq 0) \leq p_n\}.$$

It is a well-known fact that the minimax estimation rate over the class of nearly black vectors  $\ell_0[p_n]$  is known to be  $r^2(\ell_0[p_n]) = 2p_n \log(\frac{n}{p_n})(1 + o(1))$  as  $n \rightarrow \infty$  (see Donoho et al. (1992)).

If we relate this minimax rate to the oracle rate  $r^2(\theta_0)$  (given by (16)) for  $\theta_0 \in \ell_0[p_n]$ , by taking  $I_0(\theta_0) = \{i \in \mathbb{N}_n : \theta_{0,i} \neq 0\}$  we obtain trivially that

$$\sup_{\theta_0 \in \ell_0[p_n]} r^2(\theta_0) \leq \sup_{\theta_0 \in \ell_0[p_n]} r^2(I_0(\theta_0), \theta_0) \leq p_n \log\left(\frac{en}{p_n}\right) = p_n \log\left(\frac{n}{p_n}\right)(1 + o(1)).$$

Theorems 1, 2 and 4 immediately imply the adaptive minimax results for the estimation problem, the minimax contraction rate for the empirical Bayes posterior, and the coverage and size properties of the confidence set  $C(\hat{\theta}, M\hat{r})$ . We summarize these results in the following corollary.

**Corollary 2.** *Let the empirical Bayes posterior  $\hat{\pi}(\theta|X)$  be defined by (9),  $\hat{\theta}$  be defined by (12) and the confidence set  $C(\hat{\theta}, M\hat{r})$  be defined by (19). Then there exist constants  $C, c > 0$  (depending only on  $\kappa$ ) such that for any  $M > 0$ ,  $\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \|\hat{\theta} - \theta_0\|^2 \leq c p_n \log(\frac{n}{p_n})$ ,*

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \theta_0\|^2 \geq M p_n \log(\frac{n}{p_n}) | X) \leq \frac{C}{M}.$$

Moreover, for any  $t > 0$ ,  $\alpha_1, \alpha_2 \in (0, 1)$  there exist  $C_0 = C_0(\alpha_1, t)$  and  $c_0 = c_0(\alpha_2)$  such that the following relations hold

$$\sup_{\theta_0 \in \Theta_{eb}(t)} \mathbb{P}_{\theta_0}(\theta_0 \notin C(\hat{\theta}, C_0\hat{r})) \leq \alpha_1, \quad \sup_{\theta_0 \in \ell_0[p_n]} \mathbb{P}_{\theta_0}(\hat{r}^2 \geq c_0 p_n \log(\frac{n}{p_n})) \leq \alpha_2.$$

Recall that the effective size of the confidence set (19) contains an extra log factor as compared to the minimax rate  $r^2(\ell_0[p_n])$ .

The next theorem asserts in a way that the empirical Bayes posterior  $\hat{\pi}(\mathcal{I} = I|X)$  provides some “over-dimensionality” (or “undersmoothing”) control from the  $\mathbb{P}_{\theta_0}$ -perspective.

**Theorem 5.** *Let the empirical Bayes posterior  $\hat{\pi}(\mathcal{I} = I|X)$  be defined by (11). Then there exists a constant  $M > 0$  such that, for any  $1 \leq p_n \leq n$ ,*

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \hat{\pi}(|\mathcal{I}| > M p_n | X) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

**Weak  $\ell_s$ -balls** The class of *weak  $\ell_s$ -balls* for  $s \in (0, 2)$  is defined as follows:

$$m_s[p_n] = \left\{ \theta \in \mathbb{R}^n : \max_{1 \leq i \leq n} \frac{i|\theta_{[i]}|^s}{n} \leq \left(\frac{p_n}{n}\right)^s \right\}, \quad p_n = o(n) \text{ as } n \rightarrow \infty,$$

where  $|\theta_{[1]}| \geq \dots \geq |\theta_{[n]}|$  are the ordered values of  $(|\theta_1|, \dots, |\theta_n|)$ . The minimax estimation rate over this class is  $r^2(m_s[p_n]) = n\left(\frac{p_n}{n}\right)^s \left(\log\left(\frac{n}{p_n}\right)\right)^{(2-s)/2} (1+o(1))$  as  $n \rightarrow \infty$  (see Donoho and Johnstone (1994b)). Then, with  $p_n^* = n\left(\frac{p_n}{n}\right)^s \left(\log\left(\frac{n}{p_n}\right)\right)^{-s/2}$ , we derive for some  $K = K(s) > 0$  that

$$\begin{aligned} \sup_{\theta_0 \in m_s[p_n]} r^2(\theta_0) &\leq p_n^* \log\left(\frac{en}{p_n}\right) + p_n^2 n^{(2-2s)/s} \sum_{i=p_n^*+1}^{\infty} i^{-2/s} \\ &\leq p_n^* \log\left(\frac{en}{p_n}\right) + \frac{s}{2-s} p_n^2 n^{(2-2s)/s} (p_n^*)^{1-2/s} \leq K n^{1-s} p_n^s \left(\log\left(\frac{n}{p_n}\right)\right)^{1-s/2}. \end{aligned} \quad (20)$$

Theorems 1, 2 and 4 imply the following corollary.

**Corollary 3.** *Let the empirical Bayes posterior  $\hat{\pi}(\theta|X)$  be defined by (9),  $\hat{\theta}$  be defined by (12) and the confidence set  $C(\hat{\theta}, M\hat{r})$  be defined by (19). Then there exist constants  $C, c > 0$  (depending only on  $\kappa$ ) such that for any  $M > 0$ ,  $\sup_{\theta_0 \in m_s[p_n]} \mathbb{E}_{\theta_0} \|\hat{\theta} - \theta_0\|^2 \leq c n \left(\frac{p_n}{n}\right)^s [\log(\frac{n}{p_n})]^{(2-s)/2}$ ,*

$$\sup_{\theta_0 \in m_s[p_n]} \mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \theta_0\|^2 \geq M n^{1-s} p_n^s [\log(\frac{n}{p_n})]^{(2-s)/2} | X) \leq \frac{C}{M}.$$

Moreover, for any  $t > 0$ ,  $\alpha_1, \alpha_2 \in (0, 1)$  there exist  $C_0 = C_0(\alpha_1, t)$  and  $c_0 = c_0(\alpha_2)$  such that the following relations hold

$$\sup_{\theta_0 \in \Theta_{eb}(t)} \mathbb{P}_{\theta_0}(\theta_0 \notin C(\hat{\theta}, C_0 \hat{r})) \leq \alpha_1, \quad \sup_{\theta_0 \in m_s[p_n]} \mathbb{P}_{\theta_0}(\hat{r}^2 \geq c_0 n \left(\frac{p_n}{n}\right)^s [\log(\frac{n}{p_n})]^{\frac{2-s}{2}}) \leq \alpha_2.$$

The following theorem asserts in a way that the empirical Bayes posterior  $\hat{\pi}(\mathcal{I} = I|X)$  provides some “over-dimensionality” control.

**Theorem 6.** *Let the empirical Bayes posterior  $\hat{\pi}(\mathcal{I} = I|X)$  be defined by (11). Let  $s \in (0, 2)$  and  $p_n$  be such that  $p_n = o(n)$  and  $p_n^* \log(\frac{n}{p_n}) \rightarrow \infty$  as  $n \rightarrow \infty$ , where  $p_n^* = n\left(\frac{p_n}{n}\right)^s \left(\log\left(\frac{n}{p_n}\right)\right)^{-s/2}$ . Then there exists a constant  $M = M(\kappa, s) > 0$  such that*

$$\sup_{\theta_0 \in m_s[p_n]} \mathbb{E}_{\theta_0} \hat{\pi}(|\mathcal{I}| > M p_n^* | X) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

## 4 Further extensions, discussion and concluding remarks

### 4.1 Non-normality of the errors, robust empirical Bayes

It seems possible to improve the constants  $C_{or}$  and  $C_{set}$  in Theorems 1 and 2 by using more precise exponential inequalities for the normal distribution instead of simple Markov

type inequalities at some place in the proof. We however present a succinct proof rather than pursue the most accurate constant.

An even more important reason for not exploiting the specific normal structure of the errors in the model is that we can actually drop the normality assumption in the model (1). While we treat the errors  $\xi_i$  as being normal in our empirical Bayes analysis, we can allow the errors to be not necessarily normal. The model can then become *misspecified*, but all the results still hold under some moment assumptions. This means that we established that the proposed empirical Bayes approach is *robust*.

Namely, for all the results to hold, it is enough to assume that  $\xi_i$ 's are independent (not necessarily identically distributed) such that, for some  $\beta \in (0, 1]$  and  $B \geq A > 0$ ,

$$\mathbb{E}\xi_i = 0, \quad \text{Var}(\xi_i) = 1, \quad \mathbb{E}e^{\beta\xi_i^2/2} \leq e^A \leq \mathbb{E}e^{\beta\xi_i^2} \leq e^B, \quad i = 1, \dots, n. \quad (21)$$

If we use (21) instead of the normality of  $\xi_i$ 's, the proofs of the four technical lemmas will slightly change, leading to different constants in the lemmas. Hence the constants in the theorems will change as well, but the proofs of the theorems will remain the same. Let us outline what needs to be adjusted in the proofs of the lemmas.

As to Lemma 1, the normality is only used in the relation (23). We derive an analogue of (23) by using (21). First, we have

$$\mathbb{E} \exp \left\{ \frac{\beta X_i^2}{2\sigma^2} \right\} \leq \exp \left\{ \frac{\beta \theta_i^2}{\sigma^2} \right\} \mathbb{E} \exp \{ \beta \xi_i^2 \} \leq \exp \left\{ \frac{\beta \theta_i^2}{\sigma^2} + B \right\}.$$

Next, since  $(a + b)^2 \geq \frac{a^2}{2} - b^2$  for any  $a, b \in \mathbb{R}$ , we have that  $-X_i^2 \leq -\frac{\theta_i^2}{2} + \sigma^2 \xi_i^2$ . Using this relation and (21), we obtain

$$\mathbb{E} \exp \left\{ \frac{-\beta X_i^2}{2\sigma^2} \right\} \leq \exp \left\{ -\frac{\beta \theta_i^2}{4\sigma^2} \right\} \mathbb{E} \exp \left\{ \frac{\beta \xi_i^2}{2} \right\} \leq \exp \left\{ -\frac{\beta \theta_i^2}{4\sigma^2} + A \right\}.$$

By applying the last two displays to (22), we derive the assertion of Lemma 1 with  $h = \beta$  and the constants  $A_h = A_\beta = \frac{\beta}{4}$ ,  $B_h = B_\beta = \beta$ ,  $C_h = C_\beta = \frac{\beta}{2} + A$  and  $D_h = D_\beta = \frac{\beta}{2} - B$ .

Next, in the proof of Lemma 2 we take  $h = \beta$  and this lemma follows with some (different) constants  $c_1 = c_1(\beta, \kappa) > 2$ ,  $c_2 = c_2(\beta)$  and  $c_3 = c_3(\beta, \kappa)$ , for any  $\kappa \geq \kappa_0$ , with sufficiently large  $\kappa_0 = \kappa_0(\beta) > 0$ .

Lemma 3 holds true as well, but again with modified constants. Indeed, instead of (24), we get the relation

$$\mathbb{E}_{\theta_0} \hat{\pi}_I \leq \frac{\lambda_I^\beta}{c_{\kappa, n}} \exp \left\{ -\beta \sum_{i \in I_0 \setminus I} \frac{\theta_{0,i}^2}{2\sigma^2} + \kappa \beta |I_0| \log \left( \frac{en}{|I_0^\tau|} \right) + \frac{(\beta+A)|I_0^\tau|}{2} \log \left( \frac{en}{|I_0^\tau|} \right) \right\}.$$

The remainder of the adjusted proof proceeds along the lines of the present proof, provided  $\kappa \geq \kappa_0 = \kappa_0(\beta) > 0$  so that  $\sum_I \lambda_I^\beta < \infty$ . The assertion of Lemma 3 is then of the form: for any  $\varrho \in (0, 1)$ ,  $\kappa \geq \kappa_0$  and  $\tau \geq \tau_1$

$$\mathbb{E}_{\theta_0} \hat{\pi}(|\mathcal{I}| \leq \varrho |I_0^\tau| |X) \leq D \exp \left\{ -\alpha_1 |I_0^\tau| \log \left( \frac{en}{|I_0^\tau|} \right) \right\},$$

with some  $D = D(\kappa, \beta) > 0$ ,  $\alpha_1 = \alpha_1(\kappa, \varrho, \tau, \beta) > 0$  and  $\tau_1 = \tau_1(\kappa, \varrho, \beta) > 0$ .

Finally, under (21), Lemma 4 holds without normality of  $\xi_i$ 's, with another constant (instead of 6) in the relation of the lemma. The proof of this lemma is the same with the only difference that we take  $t = \beta$  (the constant  $\beta$  is from (21)) instead of  $t = \frac{1}{4}$ .

In some specific cases, the obtained bounds for the constants (and the resulting constants in the theorems) may not be very sharp as in many places we used elementary general inequalities.

## 4.2 Product prior

If, instead of the prior  $\pi$ , we take a prior  $\tilde{\pi} = \tilde{\pi}_{K,\kappa} = \sum_{I \in \mathcal{M}_n} \lambda_I \pi_I$  with  $\tau_i^2(I) = K\sigma^2 1\{i \in I\}$  for any fixed  $K > 0$  (we can even allow  $K = K_n \rightarrow \infty$ , but  $K_n = O(n)$ , as  $n \rightarrow \infty$ ) in (4) and  $\lambda_I = c_{\kappa,n} \exp\{-\kappa|I| \log n\}$  (with  $\kappa \geq \kappa_0$  for some  $\kappa_0 > 0$ ) in (6), then Theorems 1, 2, 3, Corollary 1 and Theorem 4 all hold with  $\log n$  instead of  $\log(\frac{en}{|I|})$  in all the expressions, including the oracle rate (16). This case was considered in detail in the first version of the arXiv-preprint of this paper. Thus, the results for the prior  $\tilde{\pi}$  are slightly weaker than the results obtained in this paper. For example, the minimax rates for the sparsity classes (Corollaries 2 and 3) follow from these weaker results only if the sparsity parameter  $p_n = O(n^\gamma)$  for  $\gamma \in [0, 1)$  as  $n \rightarrow \infty$ , otherwise we obtain only the so called near-minimax rates, with the factor  $\log n$  instead of  $\log(\frac{n}{p_n})$ .

However, there is an advantageous feature of the prior  $\tilde{\pi}$ . Namely, it is of product structure: if  $\lambda_I = c_\lambda \prod_{i \in I} \lambda_i$  with  $c_\lambda = \prod_{i=1}^n (1 + \lambda_i)^{-1}$ , then we compute  $\pi = \sum_{I \in \mathcal{M}} \lambda_I \pi_I = \bigotimes_{i=1}^n [\omega_i N(\mu_{1,i}, K\sigma^2) + (1 - \omega_i)\delta_0]$ ,  $\omega_i = \frac{\lambda_i}{1 + \lambda_i}$  ( $\omega_i = \lambda(i \in I)$  is the prior probability that the (random) set  $I$  contains  $i$ ). This leads to the product structure of the (empirical Bayes) posterior, and the computation of the (empirical) Bayes estimator can easily be done in the coordinatewise fashion. Indeed, in our case  $\lambda_I = c_{\kappa,n} n^{-\kappa|I|}$  (i.e.,  $\lambda_i = \lambda = n^{-\kappa}$ ) and some computations give the following empirical Bayes posterior

$$\tilde{\pi}(\theta|X) = \bigotimes_{i=1}^n [p_i N(X_i, \frac{K\sigma^2}{K+1}) + (1 - p_i)\delta_0], \quad p_i = \frac{1}{1 + h \exp\{-\frac{X_i^2}{2\sigma^2}\}},$$

where  $p_i = \tilde{\pi}(\theta_i \neq 0|X)$  and  $h = h_{\kappa,K} = \frac{\sqrt{K+1}}{\lambda} = n^\kappa(K+1)^{1/2}$ . The mean with respect to  $\tilde{\pi}(\theta|X)$  is readily obtained:  $\tilde{\theta} = E_{\tilde{\pi}}(\theta|X) = (p_i X_i, i \in \mathbb{N}_n)$ , a shrinkage estimator with easily computable shrinkage factors  $p_i$ . Coordinatewise empirical Bayes medians can also be easily computed.

## 4.3 Cardinality dependent prior $\lambda_I$

Notice that the prior  $\lambda_I$ ,  $I \in \mathcal{M}$  defined by (6) depends on the set  $I \in \mathcal{M}_n$  only via its cardinality  $|I|$ , i.e.,  $\lambda_I = g(|I|)$  for some nonnegative function  $g(k)$ ,  $k = 0, 1, \dots, n$ , such that  $\sum_{k=0}^n g(k) = 1$ .

It is easy to see that in this case  $\pi_n(k) = g(k) \binom{n}{k}$ ,  $k = 0, 1, \dots, n$ , determines the prior on the cardinality of  $I$ . Hence, the prior  $\lambda_I$  can always be modeled in two steps: first draw the the random cardinality  $K$  according to the prior  $\pi_n(k)$ , and then given  $K = k$ , draw a random set  $\mathcal{I}$  uniformly from the family of all subsets of  $\mathcal{M}_n$  of cardinality  $k$ . Castillo

and van der Vaart (2012) used such a prior  $\lambda_I$ , where the cardinality prior  $\pi_n(k)$  must be a so called “complexity prior” (see (1.2), (2.2), (2.9) and (2.10) in Castillo and van der Vaart (2012)):  $\pi_n(k) = \exp\{-ak \log(bn/k)\}$  for some  $a, b > 0$ . Since  $e^{k \log(n/k)} \leq \binom{n}{k} \leq e^{k \log(ne/k)}$ , the resulting (cardinality dependent) prior  $\lambda_I$  must satisfy

$$\exp\{-a_1|I| \log(b_1 n/|I|)\} \leq \lambda_I \leq \exp\{-a_2|I| \log(b_2 n/|I|)\}, \quad I \in \mathcal{M}.$$

The requirement of being a complexity prior from Castillo and van der Vaart (2012) corresponds to our condition  $\kappa \geq \kappa_0$  for some  $\kappa_0 > 0$ . The parameter  $\kappa$  should not be too large as the constants  $C_{or}$  and  $C_{set}$  tend to infinity as  $\kappa \rightarrow \infty$ .

Clearly, as compared to the prior  $\tilde{\pi}$  defined in subsection 4.2, there is no product structure in the resulting prior  $\pi$ , but there is still product structure in  $\pi_I$  and in the model. This partial product structure can be used to facilitate the computation of certain functionals of the posterior measure, as is demonstrated in Section 3 of Castillo and van der Vaart (2012).

#### 4.4 Computing estimators

Note that effectively the estimator (12) is a shrinkage estimator, and the estimator (14) is a hard thresholding estimator (cf. also the variable weight penalized estimator introduced by Birgé and Massart (2001)). Indeed, the estimator (12) is  $\hat{\theta}_i = p_i X_i$  where  $p_i = \sum_{I: i \in I} \hat{\pi}(\mathcal{I} = I|X)$ , and the estimator (14) is  $\tilde{\theta}_i = X_i 1\{|X_i| \geq \hat{t}\}$ , where  $\hat{t} = |X_{(\hat{k})}|$ ,  $\hat{k}$  is the minimizer of  $\sum_{i=k+1}^n X_{(i)}^2 + k(2\kappa + 1)\sigma^2 \log(\frac{en}{k})$  and  $|X_{(1)}| \geq \dots \geq |X_{(n)}|$ . The thresholding procedure is easy to implement whereas the values  $p_i$  in the shrinkage procedure are more difficult to compute. Castillo and van der Vaart (2012) demonstrated how one can use the partial product structure (in the model, but there is no product structure in the prior) to facilitate these computations.

Other estimators can be considered, for example, the coordinatewise median with respect to  $\hat{\pi}$ , which is going to be something in between shrinkage and thresholding.

#### 4.5 Extra log factor in the radial rate

Theorem 4 claims basically that the effective radial rate of the confidence set  $C(\hat{\theta}, M\hat{r})$  for  $\theta_0 \in \mathbb{R}^n$  defined by (19) is of order  $r(\theta_0)(\log n)^{1/2}$  (although the confidence set  $C(\hat{\theta}, M\hat{r})$  is slightly “smaller” than the ball  $B(\hat{\theta}, (1 + \log n)^{1/2} M\hat{r})$ ). This extra log factor enters the size property also for the two sparsity classes treated in subsection 3.4. It is not known to us whether this extra  $\log n$  factor (as compared to the oracle rate and the minimax rates for the sparsity classes) can be removed or not. Our conjecture is that this extra log factor is unavoidable if we take  $\sup_{\theta_0 \in \mathbb{R}^n}$  in the size relation of Theorem 4 ( $\Theta_1 = \mathbb{R}^n$  in (2)), but it can be removed if we take  $\sup_{\theta_0 \in \Theta_{eb}}$  instead of  $\sup_{\theta_0 \in \mathbb{R}^n}$  in the size relation of Theorem 4 ( $\Theta_1 = \Theta_{eb}$  in (2)).

#### 4.6 Empirical Bayes posterior $\tilde{\pi}(\theta|X)$

We established all the results for the empirical Bayes posterior  $\hat{\pi}(\theta|X)$  defined by (9) and the corresponding estimator  $\hat{\theta}$  defined by (12). Actually, all the theorems also hold for



the empirical Bayes posterior  $\tilde{\pi}(\theta|X)$  and the estimator  $\check{\theta}$  defined by (14). Indeed, by the definition of  $\check{I}$  and the Markov inequality, we derive that, for any  $I, I_0 \in \mathcal{M}_n$  and any  $h \in [0, 1)$ ,

$$\mathbb{P}_{\theta_0}(\check{I} = I) \leq \mathbb{P}_{\theta_0}\left(\frac{\hat{\pi}(\mathcal{I} = I|X)}{\hat{\pi}(\mathcal{I} = I_0|X)} \geq 1\right) \leq \mathbb{E}_{\theta_0}\left[\frac{\hat{\pi}(\mathcal{I} = I|X)}{\hat{\pi}(\mathcal{I} = I_0|X)}\right]^h,$$

which yields the analogue of (22). From this point on, the proof of the properties of  $\tilde{\pi}(\theta|X)$  and  $\check{\theta}$  proceeds exactly in the same way as for  $\hat{\pi}(\theta|X)$  and  $\hat{\theta}$ , with the only difference that everywhere (in the claims and in the proofs),  $1\{\check{I} = I\}$  is substituted instead of  $\hat{\pi}(\mathcal{I} = I|X)$  and  $\mathbb{P}_{\theta_0}(\check{I} = I)$  is substituted instead of  $\mathbb{E}_{\theta_0}\hat{\pi}(\mathcal{I} = I|X)$ .

## 5 Simulations

Here we present a small simulation study according to the model (1) with  $\sigma = 1$  and  $n = 500$ . We used signals  $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,n})$  of the form  $\theta_0 = (0, \dots, 0, A, \dots, A)$ , where  $p_n = \#\{\theta_{0,i} \neq 0\}$  last coordinates of  $\theta_0$  are equal to a fixed number  $A$ . Different sparsity levels  $p_n \in \{25, 50, 100\}$  and “signal strengths”  $A \in \{3, 4, 5\}$  are considered.

The simulation study concerns the construction of confidence ball by using the empirical Bayes posterior  $\hat{\pi}(\theta|X)$  defined by (9) with parameter  $\kappa = 0.7$ . We consider a ball  $B(\check{\theta}, M\hat{r}(\check{I}))$  around  $\check{\theta}$  defined by (14) with radius  $\hat{r}(\check{I}) = \sqrt{|\check{I}| \log(en/|\check{I}|)}$ , where  $\check{I}$  is given by (13) (notice that  $\hat{r}(\check{I})$  is without extra log factor). The multiplicative factor  $M$  is intended to trade-off the size of the ball against its coverage probability. For each sparsity level  $p_n \in \{25, 50, 100\}$  and signal strength  $A \in \{3, 4, 5\}$ , we simulated 100 data vectors  $X$  of dimension  $n = 500$  from the model (1) and computed the average squared radius by  $\overline{Mr^2}(\check{I})$ . Table 1 shows the ratio of  $\overline{Mr^2}(\check{I})$  to the oracle radial rate  $r^2(\theta_0) = |I_o| \log(en/|I_o|) = p_n \log(en/p_n)$ , where  $I_o$  is defined by (16), and the frequency  $\bar{\alpha}$  of the event that confidence ball  $B(\check{\theta}, M\hat{r}(\check{I}))$  contains the signal  $\theta_0$ , respectively. Notice

Table 1: The ratio of  $\overline{Mr^2}(\check{I})$  to the oracle radial rate  $r^2(\theta_0)$ , the values of multiplicative factor  $M$ , and the frequency  $\bar{\alpha}$  of the event that the confidence ball  $B(\check{\theta}, M\hat{r}(\check{I}))$  contains the signal  $\theta_0 = (0, \dots, 0, A, \dots, A)$  (where  $p_n$  last coordinates are equal to  $A$ ) computed on 100 data vectors  $X$  of length  $n = 500$  simulated from the model (1) with  $\sigma = 1$ .

$p_n$	<b>25</b>			<b>50</b>			<b>100</b>		
$A$	<b>3</b>	<b>4</b>	<b>5</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>3</b>	<b>4</b>	<b>5</b>
$\frac{\overline{Mr^2}(\check{I})}{r^2(\theta_0)}$	1.5	1.23	1.11	1.34	1.22	1.15	1.3	1.2	1.16
$M$	2.2	1.19	1	1.52	1.1	1	1.23	1.03	1
$\bar{\alpha}$	0.96	0.98	0.99	0.96	0.98	0.98	0.98	0.97	0.99

that the higher the signal strength, the closer the ratio  $\frac{\overline{Mr^2}(\check{I})}{r^2(\theta_0)}$  to 1.

## 6 Technical lemmas

First we provide a couple of technical lemmas used in the proofs of the main results.

**Lemma 1.** *Let  $\hat{\pi}_I = \hat{\pi}(\mathcal{I} = I|X)$  be defined by (11). Then the following inequality holds for any  $I, I_0 \in \mathcal{M}_n$  and  $h \in [0, 1]$ :*

$$\mathbb{E}_{\theta_0} \hat{\pi}_I \leq \left[ \frac{\lambda_I}{\lambda_{I_0}} \right]^h \exp \left\{ B_h \sum_{i \in I \setminus I_0} \frac{\theta_{0,i}^2}{\sigma^2} - A_h \sum_{i \in I_0 \setminus I} \frac{\theta_{0,i}^2}{\sigma^2} + C_h |I_0| \log \left( \frac{en}{|I_0|} \right) - D_h |I| \log \left( \frac{en}{|I|} \right) \right\},$$

where  $A_h = \frac{h}{2(1+h)}$ ,  $B_h = \frac{h}{2(1-h)}$ ,  $C_h = \frac{h}{2}$  and  $D_h = \frac{h}{2} + \frac{1}{2} \log(1-h)$ .

*Proof of Lemma 1.* According to (11), we evaluate for any  $h \in [0, 1]$  and any  $I, I_0 \in \mathcal{M}_n$

$$\begin{aligned} \mathbb{E}_{\theta_0} \hat{\pi}_I &= \mathbb{E}_{\theta_0} \frac{\lambda_I \prod_{i=1}^n \phi(X_i, X_i 1\{i \in I\}, \sigma^2 + \tau_i^2(I))}{\sum_{J \in \mathcal{M}_n} \lambda_J \prod_{i=1}^n \phi(X_i, X_i 1\{i \in J\}, \sigma^2 + \tau_i^2(J))} \\ &\leq \mathbb{E}_{\theta_0} \left[ \frac{\lambda_I \prod_{i=1}^n \phi(X_i 1\{i \notin I\}, 0, \sigma^2 + K_n(I) \sigma^2 1\{i \in I\})}{\lambda_{I_0} \prod_{i=1}^n \phi(X_i 1\{i \notin I_0\}, 0, \sigma^2 + K_n(I_0) \sigma^2 1\{i \in I_0\})} \right]^h \\ &= \left[ \frac{\lambda_I}{\lambda_{I_0}} \right]^h \mathbb{E}_{\theta_0} \exp \left\{ \frac{h}{2} \left[ \sum_{i \in I \setminus I_0} \frac{X_i^2}{\sigma^2} - \sum_{i \in I_0 \setminus I} \frac{X_i^2}{\sigma^2} - |I| \log \left( \frac{en}{|I|} \right) + |I_0| \log \left( \frac{en}{|I_0|} \right) \right] \right\}. \end{aligned} \quad (22)$$

Recall the elementary identity for  $Y \sim N(\mu_y, \sigma_y^2)$  and  $a < \sigma_y^{-2}$ :

$$\mathbb{E} \exp \left\{ \frac{aY^2}{2} \right\} = \exp \left\{ \frac{a\mu_y^2}{2(1-a\sigma_y^2)} - \frac{1}{2} \log(1 - a\sigma_y^2) \right\}. \quad (23)$$

By applying (23), the fact that  $b|I \setminus I_0| - a|I_0 \setminus I| \leq b|I| - a|I_0| \leq b|I| \log(ne/|I|)$  if  $b \geq a \geq 0$  and  $|I| \geq 1$ , we obtain the claim of the lemma with  $A_h = \frac{h}{2(1+h)}$ ,  $B_h = \frac{h}{2(1-h)}$ ,  $C_h = \frac{h}{2}$  and  $D_h = \frac{h}{2} + \frac{1}{2} \log(1-h)$ .  $\square$

Note that above lemma holds for any set  $I_0 \in \mathcal{M}_n$ . By taking  $I_0 = I_o$  defined by (16), we obtain the following lemma.

**Lemma 2.** *Let  $\hat{\pi}_I = \hat{\pi}(\mathcal{I} = I|X)$  be defined by (11), the oracle rate  $r^2(\theta_0)$  be defined by (16) and  $\kappa > 3.27$ . Then there exist positive constants  $c_1 = c_1(\kappa) > 2$ ,  $c_2$  and  $c_3 = c_3(\kappa)$  such that*

$$\mathbb{E}_{\theta_0} \hat{\pi}_I \leq \left( \frac{ne}{|I|} \right)^{-c_1 |I|} \exp \left\{ - \frac{c_2}{\sigma^2} [r^2(I, \theta_0) - c_3 r^2(\theta_0)] \right\}.$$

*Proof of Lemma 2.* Take  $h = 0.9$ , then Lemma 1 holds with the constants  $A_{0.9} = \frac{9}{38}$ ,  $B_{0.9} = \frac{9}{2}$ ,  $C_{0.9} = \frac{9}{20}$  and  $D_{0.9} = 0.45 - 0.5 \log(10)$ . Introduce the constant  $c_1 = c_1(\kappa) = 0.9\kappa + D_{0.9} - \frac{9}{38} > 2$  as  $\kappa > 3.27$ . Since  $0.9\kappa + D_{0.9} = c_1 + \frac{9}{38}$ , the definition (6) of  $\lambda_I$  entails that

$$\begin{aligned} &\left[ \frac{\lambda_I}{\lambda_{I_0}} \right]^{\frac{9}{10}} \exp \left\{ C_{0.9} |I_0| \log \left( \frac{ne}{|I_0|} \right) - D_{0.9} |I| \log \left( \frac{en}{|I|} \right) \right\} \\ &= \exp \left\{ (0.9\kappa + C_{0.9}) |I_0| \log \left( \frac{en}{|I_0|} \right) - \left( c_1 + \frac{9}{38} \right) |I| \log \left( \frac{en}{|I|} \right) \right\}. \end{aligned}$$

Using the last relation and Lemma 1 with  $h = 0.9$ , we derive that

$$\begin{aligned} \mathbb{E}_{\theta_0} \hat{\pi}_I &\leq \left\lfloor \frac{\lambda_I}{\lambda_{I_0}} \right\rfloor \frac{9}{10} e^{C_{0.9}|I_0|\log\left(\frac{en}{|I_0|}\right) - D_{0.9}|I|\log\left(\frac{en}{|I|}\right)} \exp\left\{-A_{0.9} \sum_{i \in I_0 \setminus I} \frac{\theta_{0,i}^2}{\sigma^2} + B_{0.9} \sum_{i \in I \setminus I_0} \frac{\theta_{0,i}^2}{\sigma^2}\right\} \\ &\leq \left(\frac{ne}{|I|}\right)^{-c_1|I|} \exp\left\{-\frac{9}{38} \sum_{i \in I_0 \setminus I} \frac{\theta_{0,i}^2}{\sigma^2} + E_\kappa \sum_{i \in I \setminus I_0} \frac{\theta_{0,i}^2}{\sigma^2} + E_\kappa |I_0| \log\left(\frac{en}{|I_0|}\right) - \frac{9}{38}|I| \log\left(\frac{en}{|I|}\right)\right\}, \end{aligned}$$

where  $E_\kappa = \max\left\{\frac{9}{2}, 0.9\kappa + \frac{9}{20}\right\}$ . This completes the proof, with the constants  $c_2 = \frac{9}{38}$  and  $c_3 = c_3(\kappa) = \frac{E_\kappa}{A_{0.9}} = \frac{38}{9} \max\left\{\frac{9}{2}, 0.9\kappa + \frac{9}{20}\right\}$ .  $\square$

**Lemma 3.** Let  $\hat{\pi}_I = \hat{\pi}(\mathcal{I} = I|X)$  be given by (11) with  $\kappa > 1$ . For  $\varrho \in (0, 1)$ , denote  $H_\varrho = 1 - \varrho(1 + \log(1/\varrho))$ . Then for any  $\varrho \in (0, 1)$ , any  $\tau > \frac{4\kappa(1+\varrho)+2}{H_\varrho}$  and any  $\theta_0 \in \mathbb{R}^n$ ,

$$\mathbb{E}_{\theta_0} \hat{\pi}(|\mathcal{I}| \leq \varrho |I_o^\tau| | X) \leq (e^{\kappa-1} - 1)^{-1} \exp\left\{-\alpha |I_o^\tau| \log\left(\frac{en}{|I_o^\tau|}\right)\right\},$$

where  $\alpha = \frac{\tau H_\varrho - 4\kappa(1+\varrho) - 2}{4} > 0$  and  $I_o^\tau = I_o^\tau(\theta_0)$  is defined by (16).

*Proof of Lemma 3.* For each  $\theta_0 \in \mathbb{R}^n$  and  $I \in \mathcal{M}_n$  such that  $|I| \leq \varrho |I_o^\tau|$ , define  $I_0 = I_0(I, \theta_0) = I \cup I_o^\tau$ , where  $I_o^\tau = I_o^\tau(\theta_0)$  is given by (16). As  $|I| \leq \varrho |I_o^\tau| \leq |I_o^\tau| \leq |I_0|$ , by applying (6) and (23) we obtain that

$$\begin{aligned} \mathbb{E}_{\theta_0} \hat{\pi}_I &= \mathbb{E}_{\theta_0} \frac{\lambda_I \prod_{i=1}^n \phi(X_i, X_i 1\{i \in I\}, \sigma^2 + K_n(I) \sigma^2 1\{i \in I\})}{\sum_{J \in \mathcal{M}_n} \lambda_J \prod_{i=1}^n \phi(X_i, X_i 1\{i \in J\}, \sigma^2 + K_n(J) \sigma^2 1\{i \in J\})} \\ &\leq \mathbb{E}_{\theta_0} \frac{\lambda_I}{\lambda_{I_0}} \exp\left\{-\sum_{i \in I_0 \setminus I} \frac{X_i^2}{2\sigma^2} + \frac{|I_0|}{2} \log\left(\frac{en}{|I_0|}\right) - \frac{|I|}{2} \log\left(\frac{en}{|I|}\right)\right\} \\ &\leq \frac{\lambda_I}{\lambda_{I_0}} \exp\left\{-\sum_{i \in I_0 \setminus I} \frac{\theta_{0,i}^2}{4\sigma^2} - \frac{|I_0 \setminus I|}{2} \log 2 + \frac{|I_0|}{2} \log\left(\frac{en}{|I_o^\tau|}\right) - \frac{|I|}{2} \log\left(\frac{en}{|\varrho |I_o^\tau|}\right)\right\} \\ &\leq \frac{\lambda_I}{c_{\kappa,n}} \exp\left\{-\sum_{i \in I_0 \setminus I} \frac{\theta_{0,i}^2}{4\sigma^2} + |I_0| \kappa \log\left(\frac{en}{|I_o^\tau|}\right) + \frac{|I_o^\tau|}{2} \log\left(\frac{en}{|I_o^\tau|}\right)\right\}. \end{aligned} \quad (24)$$

As the function  $x \log(ne/x)$  is increasing on the interval  $(0, n]$ , we have that  $|I| \log\left(\frac{en}{|I|}\right) \leq \varrho |I_o^\tau| \log\left(\frac{en}{|\varrho |I_o^\tau|}\right) = \varrho |I_o^\tau| (\log\left(\frac{en}{|I_o^\tau|}\right) + \log(1/\varrho)) \leq \varrho (1 + \log(1/\varrho)) |I_o^\tau| \log\left(\frac{en}{|I_o^\tau|}\right)$  for  $|I| \leq \varrho |I_o^\tau|$  with  $\varrho \in (0, 1)$ . Using this, the definition of  $I_0$  and the definition (16) of  $I_o^\tau$ , we derive that, for  $|I| \leq \varrho |I_o^\tau|$ ,

$$\begin{aligned} \frac{1}{4\sigma^2} \sum_{i \in I_0 \setminus I} \theta_{0,i}^2 &\geq \frac{1}{4\sigma^2} \left( \sum_{i \notin I} \theta_{0,i}^2 - \sum_{i \notin I_o^\tau} \theta_{0,i}^2 \right) \geq \frac{\tau}{4} (|I_o^\tau| \log\left(\frac{en}{|I_o^\tau|}\right) - |I| \log\left(\frac{en}{|I|}\right)) \\ &\geq \frac{\tau}{4} (1 - \varrho [1 + \log(1/\varrho)]) |I_o^\tau| \log\left(\frac{en}{|I_o^\tau|}\right) = \frac{\tau H_\varrho}{4} |I_o^\tau| \log\left(\frac{en}{|I_o^\tau|}\right). \end{aligned} \quad (25)$$

If  $|I| \leq \varrho |I_o^\tau|$ , then  $|I_0| \leq |I| + |I_o^\tau| \leq (1 + \varrho) |I_o^\tau|$ . This relation, (24), (25) and the facts that  $\sum_I \lambda_I = 1$  and  $c_{\kappa,n} \geq e^{\kappa-1} - 1$  imply that

$$\begin{aligned}
& \mathbb{E}_{\theta_0} \hat{\pi}(|\mathcal{I}| \leq \varrho |I_o^\tau| | X) \\
& \leq \sum_{I: |I| \leq \varrho |I_o^\tau|} \frac{\lambda_I}{c_{\kappa,n}} \exp \left\{ - \sum_{i \in I_0 \setminus I} \frac{\theta_{0,i}^2}{4\sigma^2} + |I_0| \kappa \log \left( \frac{en}{|I_o^\tau|} \right) + \frac{|I_o^\tau|}{2} \log \left( \frac{en}{|I_o^\tau|} \right) \right\} \\
& \leq c_{\kappa,n}^{-1} \sum_{I: |I| \leq \varrho |I_o^\tau|} \lambda_I \exp \left\{ - \left( \frac{\tau H_\varrho}{4} - (1 + \varrho) \kappa - \frac{1}{2} \right) |I_o^\tau| \log \left( \frac{en}{|I_o^\tau|} \right) \right\} \\
& \leq (e^{\kappa-1} - 1)^{-1} \exp \left\{ - \alpha |I_o^\tau| \log \left( \frac{en}{|I_o^\tau|} \right) \right\},
\end{aligned}$$

where  $\alpha = \frac{1}{4}(\tau H_\varrho - 4\kappa(1 + \varrho) - 2) > 0$ .  $\square$

**Lemma 4.** Let  $\xi_1, \dots, \xi_n \sim N(0, 1)$  and  $\xi_{(1)}^2 \geq \xi_{(2)}^2 \geq \dots \geq \xi_{(n)}^2$ . Then for any  $k = 1, \dots, n$

$$\mathbb{E} \sum_{i=1}^k \xi_{(i)}^2 \leq 6k \log \left( \frac{en}{k} \right).$$

*Proof.* By Jensen's inequality, we derive for any  $t < \frac{1}{2}$  that

$$\exp \left\{ t \mathbb{E} \left[ \sum_{i=1}^k \xi_{(i)}^2 \right] \right\} \leq \mathbb{E} \exp \left\{ t \sum_{i=1}^k \xi_{(i)}^2 \right\} \leq \binom{n}{k} \mathbb{E} \exp \left\{ t \sum_{i=1}^k \xi_i^2 \right\} = \frac{\binom{n}{k}}{(1-2t)^{\frac{k}{2}}}.$$

Hence  $\mathbb{E} \left[ \sum_{i=1}^k \xi_{(i)}^2 \right] \leq \frac{\log \binom{n}{k}}{t} - \frac{k \log(1-2t)}{2t}$ . Taking  $t = \frac{1}{4}$  and using  $\binom{n}{k} \leq e^{k \log(en/k)}$ , we obtain that

$$\mathbb{E} \left[ \sum_{i=1}^k \xi_{(i)}^2 \right] \leq 4k \log \left( \frac{en}{k} \right) + 2k \log 2 \leq 6k \log \left( \frac{en}{k} \right). \quad \square$$

## 7 Proofs of the theorems

Here we gather the proofs of all theorems. For a  $\tau_0 > 0$  and  $\theta_0 \in \mathbb{R}^n$ , introduce the families of sets: with the oracle rate  $r(\theta_0)$  given by (16),

$$\mathcal{O}(\tau_0) = \mathcal{O}(\tau_0, \theta_0) = \{I \in \mathcal{M}_n : r^2(I, \theta_0) \leq \tau_0 r^2(\theta_0)\}, \quad (26)$$

$$\mathcal{O}^c(\tau_0) = \mathcal{M}_n \setminus \mathcal{O}(\tau_0) = \{I \in \mathcal{M}_n : r^2(I, \theta_0) > \tau_0 r^2(\theta_0)\}. \quad (27)$$

The families  $\mathcal{O}(\tau_0)$  and  $\mathcal{O}^c(\tau_0)$  form a partition of  $\mathcal{M}_n$ :  $\mathcal{M}_n = \mathcal{O}(\tau_0) \sqcup \mathcal{O}^c(\tau_0)$ .

*Proof of Theorem 1.* Recall that, according to (10),  $\hat{\pi}(\theta | X, \mathcal{I} = I) = \bigotimes_{i \in \mathbb{N}_n} N(X_i 1\{i \in I\}, \frac{K_n(I) \sigma^2 1\{i \in I\}}{K_n(I)+1})$ . Let  $\hat{\mathbb{E}}$  and  $\widehat{\text{var}}$  denote the (random) expectation and variance with

respect to  $\hat{\pi}(\theta|X, \mathcal{I} = I)$ , then

$$\begin{aligned}\hat{\mathbb{E}}(\|\theta - \theta_0\|^2|X, \mathcal{I} = I) &= \sum_{i \in \mathbb{N}_n} \widehat{\text{var}}(\theta_i|X, \mathcal{I} = I) + \sum_{i \in \mathbb{N}_n} (\hat{\mathbb{E}}(\theta_i|X, \mathcal{I} = I) - \theta_{0,i})^2 \\ &= \frac{K_n(I)\sigma^2|I|}{K_n(I) + 1} + \sigma^2 \sum_{i \in I} \xi_i^2 + \sum_{i \in I^c} \theta_{0,i}^2 \leq r^2(I, \theta_0) + \sigma^2 \sum_{i \in I} \xi_i^2,\end{aligned}$$

where  $\xi_i = \sigma^{-1}(X_i - \theta_{0,i}) \sim N(0, 1)$ . The last relation and the Markov inequality imply that

$$\begin{aligned}\mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \theta_0\| \geq Mr(\theta_0)|X) &= \mathbb{E}_{\theta_0} \sum_{I \in \mathcal{M}_n} \hat{\pi}(\|\theta - \theta_0\| \geq Mr(\theta_0)|X, \mathcal{I} = I) \hat{\pi}_I \\ &\leq \mathbb{E}_{\theta_0} \sum_{I \in \mathcal{M}_n} \frac{\hat{\mathbb{E}}(\|\theta - \theta_0\|^2|X, \mathcal{I} = I)}{M^2 r^2(\theta_0)} \hat{\pi}_I \\ &\leq \frac{\sum_{I \in \mathcal{M}_n} r^2(I, \theta_0) \mathbb{E}_{\theta_0} \hat{\pi}_I}{M^2 r^2(\theta_0)} + \frac{\mathbb{E}_{\theta_0} \left[ \sum_{I \in \mathcal{M}_n} (\sigma^2 \sum_{i \in I} \xi_i^2) \hat{\pi}_I \right]}{M^2 r^2(\theta_0)}.\end{aligned}\quad (28)$$

Let the sets  $\mathcal{O}(\tau_0)$  and  $\mathcal{O}^c(\tau_0)$  be defined by (26) and (27), respectively. Let  $\tau_0$  be chosen in such a way that  $\tau_0 > c_3$ , where  $c_3 = \frac{38}{9} \max\{\frac{9}{2}, 0.9\kappa + \frac{9}{20}\}$  is defined in the proof of Lemma 2 and  $\kappa > 3.27$ . For  $I \in \mathcal{O}^c(\tau_0)$ , we evaluate

$$r^2(I, \theta_0) - c_3 r^2(\theta_0) \geq (1 - \frac{c_3}{\tau_0}) r^2(I, \theta_0). \quad (29)$$

Denote  $B = B(\kappa, \tau_0) = \frac{c_2(\tau_0 - c_3)}{2\tau_0} = \frac{9(\tau_0 - c_3)}{76\tau_0}$ , where  $c_2 = \frac{9}{38}$  is defined in the proof of Lemma 2. Using Lemma 2, (29) and the facts that  $\max_{x \geq 0} \{xe^{-cx}\} \leq (ce)^{-1}$  (for any  $c > 0$ ) and  $\binom{n}{k} \leq (\frac{ne}{k})^k$ , we obtain that

$$\begin{aligned}&\sum_{I \in \mathcal{O}^c(\tau_0)} r^2(I, \theta_0) [\mathbb{E}_{\theta_0} \hat{\pi}_I]^{\frac{1}{2}} \\ &\leq \sum_{I \in \mathcal{O}^c(\tau_0)} r^2(I, \theta_0) \left(\frac{ne}{|I|}\right)^{-\frac{c_1|I|}{2}} \exp\left\{-\frac{c_2}{2\sigma^2} [r^2(I, \theta_0) - c_3 r^2(\theta_0)]\right\} \\ &\leq \sum_{I \in \mathcal{O}^c(\tau_0)} \left(\frac{ne}{|I|}\right)^{-\frac{c_1|I|}{2}} r^2(I, \theta_0) e^{-\frac{B}{\sigma^2} r^2(I, \theta_0)} \leq \frac{\sigma^2}{Be} \sum_{I \in \mathcal{O}^c(\tau_0)} \left(\frac{ne}{|I|}\right)^{-\frac{c_1|I|}{2}} \\ &\leq \frac{\sigma^2}{Be} \sum_{k=1}^n \binom{n}{k} \left(\frac{ne}{k}\right)^{-\frac{c_1 k}{2}} \leq \frac{\sigma^2}{Be} \sum_{k=1}^n \left(\frac{ne}{k}\right)^{-\frac{k(c_1-2)}{2}} \leq \frac{\sigma^2}{Be(e^{(c_1-2)/2} - 1)},\end{aligned}\quad (30)$$

where  $c_1 = c_1(\kappa) > 2$  is defined in the proof of Lemma 2.

Let  $|I_{max}| = \max\{|J| : J \in \mathcal{O}(\tau_0)\}$  and  $\xi_{(1)}^2 \geq \xi_{(2)}^2 \geq \dots \geq \xi_{(n)}^2$  denote the non-increasing rearrangement of  $\xi_1^2, \dots, \xi_n^2$ . From Lemma 4, it follows that

$$\begin{aligned}\mathbb{E}_{\theta_0} \left[ \sum_{I \in \mathcal{O}(\tau_0)} \hat{\pi}_I \sigma^2 \sum_{i \in I} \xi_i^2 \right] &\leq \sigma^2 \mathbb{E}_{\theta_0} \left[ \sum_{i=1}^{|I_{max}|} \xi_{(i)}^2 \right] \\ &\leq 6\sigma^2 |I_{max}| \log\left(\frac{ne}{|I_{max}|}\right) \leq 6\tau_0 r^2(\theta_0),\end{aligned}\quad (31)$$

as  $I_{max} \in \mathcal{O}(\tau_0)$ . We have  $E(\sum_{i \in I} \xi_i^2)^2 = |I|^2 + 2|I| \leq 3|I|^2$ . Using this, Cauchy-Schwarz inequality and (30), we evaluate

$$\begin{aligned} E_{\theta_0} \left[ \sum_{I \in \mathcal{O}^c(\tau_0)} \hat{\pi}_I \sigma^2 \sum_{i \in I} \xi_i^2 \right] &\leq \sum_{I \in \mathcal{O}^c(\tau_0)} \sigma^2 \left[ E_{\theta_0} \left( \sum_{i \in I} \xi_i^2 \right)^2 \right]^{\frac{1}{2}} \left[ E_{\theta_0} \hat{\pi}_I^2 \right]^{\frac{1}{2}} \\ &\leq \sqrt{3} \sum_{I \in \mathcal{O}^c(\tau_0)} r^2(I, \theta_0) \left[ E_{\theta_0} \hat{\pi}_I \right]^{\frac{1}{2}} \leq \frac{\sqrt{3} \sigma^2}{Be(e^{(c_1-2)/2} - 1)}. \end{aligned} \quad (32)$$

From (31) and (32), it follows that

$$\begin{aligned} E_{\theta_0} \left[ \sum_{I \in \mathcal{M}_n} \hat{\pi}_I \sigma^2 \sum_{i \in I} \xi_i^2 \right] &= E_{\theta_0} \left[ \sum_{I \in \mathcal{O}(\tau_0)} \hat{\pi}_I \sigma^2 \sum_{i \in I} \xi_i^2 + \sum_{I \in \mathcal{O}^c(\tau_0)} \hat{\pi}_I \sigma^2 \sum_{i \in I} \xi_i^2 \right] \\ &\leq 6\tau_0 r^2(\theta_0) + \frac{\sqrt{3} \sigma^2}{Be(e^{(c_1-2)/2} - 1)}. \end{aligned} \quad (33)$$

Recall that  $\sum_I \hat{\pi}_I = 1$  and  $r^2(I, \theta_0) \leq \tau_0 r^2(I_o, \theta_0)$  for all  $I \in \mathcal{O}(\tau_0)$ . Using these relations and (30), we have

$$\begin{aligned} \sum_{I \in \mathcal{M}_n} r^2(I, \theta_0) E_{\theta_0} \hat{\pi}_I &= \sum_{I \in \mathcal{O}(\tau_0)} r^2(I, \theta_0) E_{\theta_0} \hat{\pi}_I + \sum_{I \in \mathcal{O}^c(\tau_0)} r^2(I, \theta_0) E_{\theta_0} \hat{\pi}_I \\ &\leq \tau_0 r^2(\theta_0) + \sum_{I \in \mathcal{O}^c(\tau_0)} r^2(I, \theta_0) E_{\theta_0} \hat{\pi}_I \leq \tau_0 r^2(\theta_0) + \frac{\sigma^2}{Be(e^{(c_1-2)/2} - 1)}. \end{aligned} \quad (34)$$

Finally, combining the relations (28), (33) and (34), and taking into account that  $r^2(\theta_0) \geq \sigma^2$ , we finish the proof:

$$E_{\theta_0} \hat{\pi} \left( \|\theta - \theta_0\|^2 \geq M^2 r^2(\theta_0) |X \right) \leq \frac{7\tau_0}{M^2} + \frac{\sigma^2(\sqrt{3}+1)}{M^2 Be(e^{(c_1-2)/2} - 1) r^2(\theta_0)} \leq \frac{C_{or}}{M^2},$$

where  $C_{or} = 7\tau_0 + \frac{\sqrt{3}+1}{Be(e^{(c_1-2)/2} - 1)} = 7\tau_0 + \frac{76\tau_0(\sqrt{3}+1)}{9e(\tau_0-c_3)(e^{(c_1-2)/2} - 1)}$ , and we take, say,  $\tau_0 = c_3 + 1$ . Recall that  $c_3 = \frac{38}{9} \max \left\{ \frac{9}{2}, 0.9\kappa + \frac{9}{20} \right\}$ .  $\square$

*Proof of Theorem 2.* The proof of this theorem is essentially contained in the proof of Theorem (1). By using Fubini's theorem, Cauchy-Schwarz inequality and the relations (12), (33), (34), we obtain that

$$\begin{aligned} E_{\theta_0} \|\hat{\theta} - \theta_0\|^2 &= E_{\theta_0} \sum_{i \in \mathbb{N}_n} \left( \sum_{I \in \mathcal{M}_n} X_i(I) \hat{\pi}_I - \theta_{0,i} \right)^2 \\ &\leq E_{\theta_0} \sum_{I \in \mathcal{M}_n} \|X_i(I) - \theta_{0,i}\|^2 \hat{\pi}_I = E_{\theta_0} \sum_{I \in \mathcal{M}_n} \left( \sigma^2 \sum_{i \in I} \xi_i^2 + \sum_{i \in I^c} \theta_{0,i}^2 \right) \hat{\pi}_I \\ &\leq E_{\theta_0} \sum_{I \in \mathcal{M}_n} \left( \sigma^2 \sum_{i \in I} \xi_i^2 + r^2(I, \theta_0) \right) \hat{\pi}_I \leq C''_{est} r^2(\theta_0) + C'_{est} \sigma^2 \leq C_{est} r^2(\theta_0), \end{aligned}$$

where  $C''_{est} = 7\tau_0$ ,  $C'_{est} = \frac{76\tau_0(\sqrt{3}+1)}{9e(\tau_0-c_3)(e^{(c_1-2)/2} - 1)}$ ,  $C_{est} = C''_{est} + C'_{est} = C_{or}$ , with  $\tau_0, c_1, c_3$  and  $C_{or}$  defined in the proof of Theorem 1.  $\square$

*Proof of Theorem 3.* Recall one technical fact. Let  $\Lambda(S)$  be the Lebesgue measure (or volume) of a set  $S \subset \mathbb{R}^k$  and  $B_k(r) = \{x \in \mathbb{R}^k : \|x\| \leq r\}$  be the Euclidean ball of radius  $r$  in space  $\mathbb{R}^k$ . Then

$$\Lambda(B_k(r)) \leq e\pi^{-1/2}(2\pi)^{k/2}k^{-1/2}\left(\frac{er^2}{k}\right)^{k/2}. \quad (35)$$

The proof can be found, e.g., in Belitser (2014).

Denote  $\hat{p}_I = \hat{\pi}(\|\theta - \tilde{\theta}\|^2 \leq \delta^2 \sigma^2 |I_o^\tau| |X, \mathcal{I} = I)$ , where  $I_o^\tau = I_o^\tau(\theta_0)$  is defined by (16). Recall that, according to (10),

$$\hat{\pi}(\theta | X, \mathcal{I} = I) = \bigotimes_{i \in \mathbb{N}_n} N(X_i 1\{i \in I\}, \frac{K_n(I)\sigma^2 1\{i \in I\}}{K_n(I)+1}).$$

Let  $Z_1, \dots, Z_n$  be independent  $N(0, 1)$ . Using Anderson's inequality and (35), we derive that, with  $P_{\theta_0}$ -probability 1, for any  $\delta \in [0, 1]$ ,

$$\begin{aligned} \hat{p}_I &= \hat{\pi}\left(\sum_{i \in I} (X_i + \sqrt{\frac{K_n(I)}{K_n(I)+1}} \sigma Z_i - \tilde{\theta}_i(X))^2 + \sum_{i \notin I} \tilde{\theta}_i^2 \leq \delta^2 \sigma^2 |I_o^\tau| |X, \mathcal{I} = I\right) \\ &\leq P\left\{\frac{(ne-|I|)}{ne} \sum_{i \in I} Z_i^2 \leq \delta^2 |I_o^\tau|\right\} \leq P\left\{\sum_{i \in I} Z_i^2 \leq \frac{\delta^2 e |I_o^\tau|}{e-1}\right\} \\ &\leq \frac{\Lambda\left(B_{|I|}\left(\delta \sqrt{\frac{e |I_o^\tau|}{e-1}}\right)\right)}{(2\pi)^{|I|/2}} \leq \frac{e}{\sqrt{\pi |I|}} \left(\frac{e^2 \delta^2 |I_o^\tau|}{|I|(e-1)}\right)^{|I|/2}. \end{aligned} \quad (36)$$

Recall the constants  $\varrho$ ,  $\alpha$  and  $\tau$  from Lemma 3. We fix  $\varrho = e^{-1}$  so that  $\alpha = \frac{\tau(1-\varrho[1+\log(1/\varrho)])-4\kappa(1+\varrho)-2}{4e} = \frac{\tau(e-2)-4\kappa(e+1)-2e}{4e}$ . To apply Lemma 3,  $\tau$  must satisfy  $\tau > \frac{4\kappa(1+\varrho)+2}{1-\varrho[1+\log(1/\varrho)]} = \frac{4\kappa(e+1)+2e}{e-2}$ , which is the condition of the theorem. Consider two cases  $e^{-\alpha |I_o^\tau| \log(en/|I_o^\tau|)} \leq \delta$  and  $e^{-\alpha |I_o^\tau| \log(en/|I_o^\tau|)} > \delta$ , where  $I_o^\tau = I_o^\tau(\theta_0)$  is defined by (16).

First suppose  $e^{-\alpha |I_o^\tau| \log(en/|I_o^\tau|)} \leq \delta$ . Applying (36) and Lemma 3 with  $\varrho = e^{-1}$ , we obtain that, for  $e^{-\alpha |I_o^\tau| \log(en/|I_o^\tau|)} \leq \delta$ ,

$$\begin{aligned} E_{\theta_0} \hat{\pi}(\|\theta - \tilde{\theta}\|^2 \leq \delta^2 \sigma^2 |I_o^\tau| |X) &= E_{\theta_0} \sum_{I \in \mathcal{M}_n} \hat{p}_I \hat{\pi}_I \\ &\leq \sum_{I: |I| \geq \varrho |I_o^\tau|} \frac{e}{\sqrt{\pi |I|}} \left(\frac{e^2 \delta^2 |I_o^\tau|}{|I|(e-1)}\right)^{|I|/2} E_{\theta_0} \hat{\pi}_I + E_{\theta_0} \hat{\pi}(|\mathcal{I}| < \varrho |I_o^\tau| |X) \\ &\leq C_1 \delta \sum_{I \in \mathcal{M}_n} \frac{1}{\sqrt{|I|}} \left(\frac{e^2 \delta^2}{\varrho(e-1)}\right)^{(|I|-1)/2} E_{\theta_0} \hat{\pi}_I + \frac{e^{-\alpha |I_o^\tau| \log(ne/|I_o^\tau|)}}{e^{\kappa-1} - 1} \\ &\leq (C_1 + (e^{\kappa-1} - 1)^{-1}) \delta, \end{aligned} \quad (37)$$

if  $\frac{e^2 \delta^2}{\varrho(e-1)} \leq 1$ , or  $\delta \leq \frac{\sqrt{\varrho(e-1)}}{e} = \frac{(e-1)^{1/2}}{e^{3/2}}$ . Here  $C_1 = \frac{e^2}{\sqrt{\pi \varrho(e-1)}} = \frac{e^{5/2}}{\sqrt{\pi(e-1)}}$ .

Now consider the case  $e^{-\alpha|I_o^\tau|\log(ne/|I_o^\tau|)} > \delta$ . Then  $|I_o^\tau| < \frac{\log(\delta^{-1})}{\alpha \log(ne/|I_o^\tau|)}$ . By using this and (36), we derive

$$\begin{aligned}
\mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \tilde{\theta}\|^2 \leq \delta^2 \sigma^2 |I_o^\tau| |X) &= \mathbb{E}_{\theta_0} \sum_{I \in \mathcal{M}_n} \hat{p}_I \hat{\pi}_I \\
&\leq \sum_{I \in \mathcal{M}_n} \frac{e}{\sqrt{\pi|I|}} \left( \frac{e^2 \delta^2 |I_o^\tau|}{|I|(e-1)} \right)^{|I|/2} \mathbb{E}_{\theta_0} \hat{\pi}_I \\
&\leq C_3 \delta \left( \frac{\log(\delta^{-1})}{\log(ne/|I_o^\tau|)} \right)^{1/2} \sum_{I \in \mathcal{M}_n} \frac{(C_2 \delta^2 \frac{\log(\delta^{-1})}{\log(ne/|I_o^\tau|)})^{(|I|-1)/2}}{|I|^{(|I|+1)/2}} \mathbb{E}_{\theta_0} \hat{\pi}_I \\
&\leq C_4 \delta \left( \frac{\log(\delta^{-1})}{\log(ne/|I_o^\tau|)} \right)^{1/2} \leq C_4 \delta [\log(\delta^{-1})]^{1/2},
\end{aligned} \tag{38}$$

where  $C_2 = \frac{e^2}{\alpha(e-1)}$ ,  $C_3 = e(C_2/\pi)^{1/2}$  and  $C_4 = C_3 \max_{k \in \mathbb{N}} \frac{C_5^{(k-1)/2}}{k^{(k+1)/2}}$ , with  $C_5 = C_2 \max_{\delta \in [0,1]} \delta^2 \log(\delta^{-1}) = C_2/(2e)$ ,  $\alpha = \frac{\tau(e-2)-4\kappa(e+1)-2e}{4e}$ .

The relation (37) holds if  $e^{-\alpha|I_o^\tau|\log(ne/|I_o^\tau|)} \leq \delta \leq \frac{(e-1)^{1/2}}{e^{3/2}} = \bar{C}_2 < 1$  and the relation (38) holds if  $\delta < e^{-\alpha|I_o^\tau|\log(ne/|I_o^\tau|)}$ . Combining these two concludes the proof of the theorem: for  $\bar{C}_1 = \max \{C_4, C_1 + (e^{\kappa-1} - 1)^{-1}\}$  and any  $\delta \in (0, \bar{C}_2]$ ,

$$\mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \tilde{\theta}\|^2 \leq \delta^2 \sigma^2 |I_o^\tau| |X) \leq \bar{C}_1 \delta [\log(\delta^{-1})]^{1/2}. \quad \square$$

*Proof of Theorem 4.* By the Markov inequality and Theorem 2,

$$\sup_{\theta_0 \in \mathbb{R}^n} \mathbb{P}_{\theta_0}(\|\theta_0 - \hat{\theta}\| \geq Mr(\theta_0)) \leq \frac{\mathbb{E}_{\theta_0} \|\theta_0 - \hat{\theta}\|^2}{M^2 r^2(\theta_0)} \leq \frac{C_{est}}{M^2}. \tag{39}$$

Since  $r^2(\theta_0) \leq r_\tau^2(\theta_0)$  for  $\tau \geq 1$ , Corollary 1 yields that for any  $\theta_0 \in \mathbb{R}^n$ , any estimator  $\tilde{\theta} = \tilde{\theta}(X)$  and any  $\delta \in \left(0, \frac{[\log(en/|I_o^\tau|)]^{-1/2} \bar{C}_2}{t_\tau(\theta_0) + \tau}\right]$ ,

$$\mathbb{E}_{\theta_0} \hat{\pi}(\|\theta - \tilde{\theta}\| \leq \delta r(\theta_0) |X) \leq \bar{C}_1 [(t_\tau(\theta_0) + \tau) \log(\frac{en}{|I_o^\tau|})]^{1/2} \delta [\log(\delta^{-1})]^{1/2},$$

First we bound the coverage probability of the confidence set (19). By using (18) with  $\gamma = \frac{1}{2}$  and the Markov inequality, we obtain

$$\begin{aligned}
&\mathbb{P}_{\theta_0}(\theta_0 \notin C(\hat{\theta}, M\hat{r})) \\
&\leq \mathbb{P}_{\theta_0}(\|\theta_0 - \hat{\theta}\| > M(\log \frac{en}{|I_o^\tau|})^{1/2} \hat{r}, \hat{r} \geq \delta r(\theta_0)) + \mathbb{P}_{\theta_0}(\hat{r} < \delta r(\theta_0)) \\
&\leq \mathbb{P}_{\theta_0}(\|\theta_0 - \hat{\theta}\| > M\delta(\log \frac{en}{|I_o^\tau|})^{1/2} r(\theta_0)) + \mathbb{P}_{\theta_0}(\hat{\pi}(\|\theta - \hat{\theta}\| \leq \delta r(\theta_0) |X) \geq \frac{1}{2}) \\
&\leq \mathbb{P}_{\theta_0}(\|\theta_0 - \hat{\theta}\| > M\delta(\log \frac{en}{|I_o^\tau|})^{1/2} r(\theta_0)) + 2\mathbb{E}_{\theta_0}(\hat{\pi}(\|\theta - \hat{\theta}\| \leq \delta r(\theta_0) |X)).
\end{aligned}$$

The last two displays and (39) imply that for each  $\theta_0 \in \Theta_{eb}(t)$

$$\mathbb{P}_{\theta_0}(\theta_0 \notin C(\hat{\theta}, M\hat{r})) \leq \frac{C_{est}}{M^2 \delta^2 \log(\frac{en}{|I_o^\tau|})} + 2\bar{C}_1 \delta [\log(\frac{en}{|I_o^\tau|}) \log(\delta^{-1})]^{1/2}, \tag{40}$$



for any  $M > 0$  and any  $\delta \in (0, [\log(\frac{en}{|I_\delta^c|})]^{-1/2}\tilde{C}_2]$ , where  $\tilde{C}_1 = \tilde{C}_1(t) = \bar{C}_1\sqrt{t+\tau}$ ,  $\tilde{C}_2 = \tilde{C}_2(t) = \frac{\bar{C}_2}{\sqrt{t+\tau}}$ , and  $\bar{C}_1, \bar{C}_2$  are defined in Corollary 1. For  $\alpha_1 \in (0, 1)$ , define

$$\delta_1 = \max \left\{ \delta \in (0, [\log(\frac{en}{|I_\delta^c|})]^{-1/2}\tilde{C}_2] : 2\tilde{C}_1\delta [\log(\frac{en}{|I_\delta^c|})\log(\delta^{-1})]^{1/2} \leq \frac{\alpha_1}{2} \right\}.$$

Clearly,  $\delta_1 \geq [\log(\frac{en}{|I_\delta^c|})]^{-1/2}c_0$  for some  $c_0 = c_0(\alpha_1, t)$ . Next,

$$\min \left\{ M \in \mathbb{R} : \frac{C_{est}}{M^2\delta_1^2\log(\frac{en}{|I_\delta^c|})} \leq \frac{\alpha_1}{2} \right\} = \sqrt{\frac{2C_{est}}{\alpha_1\delta_1^2\log(\frac{en}{|I_\delta^c|})}} \leq \sqrt{\frac{2C_{est}}{\alpha_1c_0^2}} = M_0,$$

$M_0 = M_0(\alpha_1, t)$ . Taking  $\delta = \delta_1$  in (40), we obtain that for all  $M \geq M_0$ ,

$$\sup_{\theta_0 \in \Theta_{eb}(t)} \mathbb{P}_{\theta_0}(\theta_0 \notin C(\hat{\theta}, M\hat{r})) \leq \frac{\alpha_1}{2} + \frac{\alpha_1}{2} = \alpha_1. \quad (41)$$

Now we ensure the size property. By the conditional Markov inequality, (18) with  $\gamma = \frac{1}{2}$ , (19), (39) and Theorem 1, we derive that, for all  $\theta_0 \in \mathbb{R}^n$ ,

$$\begin{aligned} \mathbb{P}_{\theta_0}(\hat{r} \geq Mr(\theta_0)) &\leq \mathbb{P}_{\theta_0}(\hat{\pi}(\|\theta - \hat{\theta}\| \leq Mr(\theta_0)|X) \leq \tfrac{1}{2}) \\ &= \mathbb{P}_{\theta_0}(\hat{\pi}(\|\theta - \hat{\theta}\| > Mr(\theta_0)|X) > \tfrac{1}{2}) \leq 2\mathbb{E}_{\theta_0}(\hat{\pi}(\|\theta - \hat{\theta}\| > Mr(\theta_0)|X)) \\ &\leq 2\mathbb{E}_{\theta_0}[\hat{\pi}(\|\theta - \theta_0\| \geq \tfrac{Mr(\theta_0)}{2}|X)] + 2\mathbb{E}_{\theta_0}[\hat{\pi}(\|\theta_0 - \hat{\theta}\| \geq \tfrac{Mr(\theta_0)}{2}|X)] \\ &\leq \frac{8(C_{or} + C_{est})}{M^2}. \end{aligned}$$

Thus,  $\sup_{\theta_0 \in \mathbb{R}^n} \mathbb{P}_{\theta_0}(\hat{r} \geq Mr(\theta_0)) \leq \frac{8(C_{or} + C_{est})}{M^2}$ . For  $\alpha_1 \in (0, 1)$ , take  $C_0 = \min \{M > 0 : \frac{8(C_{or} + C_{est})}{M^2} \leq \alpha_2\}$ , so that for all  $C \geq C_0 = C_0(\alpha_2)$

$$\sup_{\theta_0 \in \mathbb{R}^n} \mathbb{P}_{\theta_0}(\hat{r} \geq Cr(\theta_0)) \leq \alpha_2. \quad (42)$$

The proof is complete as we established (41) and (42).  $\square$

*Proof of Theorem 5.* For each  $\theta_0 \in \ell_0[p_n]$ , the oracle rate  $r^2(\theta_0) \leq p_n \log(\frac{en}{p_n})$ . On the other hand, if  $|I| > Mp_n$ , then  $r^2(I, \theta_0) \geq |I| \log(\frac{en}{|I|}) \geq Mp_n \log(\frac{en}{Mp_n})$ . Thus, for any  $c > 0$ ,  $\theta_0 \in \ell_0[p_n]$  and  $|I| > Mp_n$ , we have that

$$r^2(I, \theta_0) - cr^2(\theta_0) \geq (M - c)p_n \log(\frac{en}{p_n}) - Mp_n \log M. \quad (43)$$

Recall the positive constants  $c_1, c_2, c_3$  defined in Lemma 2. Let  $M > c_3$ , then by using Lemma 2, (43) and the fact that  $\sum_{I \in \mathcal{M}_n} (\frac{ne}{|I|})^{-c_1|I|} \leq C$  for  $c_1 > 2$  (see (30)), we obtain that

$$\begin{aligned} \sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \hat{\pi}(|I| > Mp_n|X) &= \sup_{\theta_0 \in \ell_0[p_n]} \sum_{I \in \mathcal{M}_n: |I| > Mp_n} \mathbb{E}_{\theta_0} \hat{\pi}(I = I|X) \\ &\leq \sup_{\theta_0 \in \ell_0[p_n]} \sum_{I \in \mathcal{M}_n: |I| > Mp_n} \left(\frac{ne}{|I|}\right)^{-c_1|I|} \exp \left\{ -c_2(r^2(I, \theta_0) - c_3r^2(\theta_0)) \right\} \\ &\leq C \exp \left\{ -c_2(M - c_3)p_n \log(\frac{en}{p_n}) + c_2Mp_n \log M \right\} \rightarrow 0, \end{aligned}$$

as  $n \rightarrow \infty$ , which proves the claimed result.  $\square$

*Proof of Theorem 6.* Recall (20):  $r^2(\theta_0) \leq Kn^{1-s}p_n^s(\log(\frac{n}{p_n}))^{(2-s)/2}$  for each  $\theta_0 \in m_s[p_n]$  for some  $K = K(s)$ . On the other hand, if  $|I| > Mp_n^*$ , then  $r^2(I, \theta_0) \geq |I| \log(\frac{en}{|I|}) \geq Mp_n^* \log(\frac{en}{Mp_n^*})$  and  $\log(\frac{en}{Mp_n^*}) = s \log(\frac{n}{p_n}) + \frac{s}{2} \log \log(\frac{n}{p_n}) + \log(\frac{e}{M}) \geq s \log(\frac{n}{p_n}) - \log M$  for sufficiently large  $n$  as  $p_n = o(n)$ . Then, for any  $c > 0$ ,  $\theta_0 \in m_s[p_n]$ ,  $M > \frac{cK}{s}$  and  $|I| > Mp_n^*$ , we have that for sufficiently large  $n$

$$\begin{aligned} r^2(I, \theta_0) - cr^2(\theta_0) &\geq Mp_n^* \log(\frac{en}{Mp_n^*}) - cKn(p_n/n)^s(\log(\frac{n}{p_n}))^{(2-s)/2} \\ &= Mp_n^* \log(\frac{en}{Mp_n^*}) - cKp_n^* \log(\frac{n}{p_n}) \geq p_n^*(Ms - cK) \log(\frac{n}{p_n}) - Mp_n^* \log M. \end{aligned} \quad (44)$$

Recall the positive constants  $c_1 = c_1(\kappa)$ ,  $c_2$  and  $c_3 = c_3(\kappa)$  defined in Lemma 2. Let  $M > \frac{c_3K}{s}$ , then by using Lemma 2, (44) and the fact that  $\sum_{I \in \mathcal{M}_n} (\frac{ne}{|I|})^{-c_1|I|} \leq C$  (see (30)), we obtain

$$\begin{aligned} \sup_{\theta_0 \in m_s[p_n]} E_{\theta_0} \hat{\pi}(|\mathcal{I}| > Mp_n^* | X) &= \sup_{\theta_0 \in m_s[p_n]} \sum_{I \in \mathcal{M}_n: |I| > Mp_n^*} E_{\theta_0} \hat{\pi}(\mathcal{I} = I | X) \\ &\leq \sup_{\theta_0 \in m_s[p_n]} \sum_{I \in \mathcal{M}_n: |I| > Mp_n^*} \left(\frac{ne}{|I|}\right)^{-c_1|I|} \exp \left\{ -c_2(r^2(I, \theta_0) - c_3cr^2(\theta_0)) \right\} \\ &\leq C \exp \left\{ -c_2p_n^*(Ms - c_3K) \log(\frac{n}{p_n}) + Mp_n^* \log M \right\} \rightarrow 0, \quad \text{as } n \rightarrow \infty, \end{aligned}$$

which proves the desired result.  $\square$

## References

- [1] ABRAMOVICH, F., BENJAMINI, Y. DONOHO, D.L. and JOHNSTONE, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* 34, 584–653.
- [2] ABRAMOVICH, F., GRINSHTEIN, V. and PENSKEY, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.* 35, 2261–2286.
- [3] BABENKO, A. and BELITSER, E. (2010). Oracle projection convergence rate of posterior. *Math. Meth. Statist.* 19, 219–245.
- [4] BARAUD, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* 32, 528–551.
- [5] BELITSER, E. (2014). On coverage and oracle radial rate of DDM-credible sets. arXiv:1407.5232.
- [6] BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* 3, 203–268.
- [7] BULL, A. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Statist.* 6, 1490–1516.

- [8] BULL, A. and NICKL, R. (2013). Adaptive confidence sets in  $\ell_2$ . *Probab. Theory and Rel. Fields.* 156, 889–919.
- [9] CAI, T.T. and LOW, M.G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* 32, 1805–1840.
- [10] CARVALHO, C.M., POLSON N.G. and SCOTT J.G. (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika* 97, 465–480.
- [11] CASTILLO, I. and VAN DER VAART, A.W. (2012). Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* 40, 2069–2101.
- [12] DONOHO, D.L., JOHNSTONE, I.M., HOCH, J.C. and STERN, A.S. (1992). Maximum entropy and the nearly black object(with Discussion). *J.Roy. Statist.Soc.Ser. B* 54, 41–81.
- [13] DONOHO, D.L. and JOHNSTONE, I.M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425–455.
- [14] DONOHO, D.L. and JOHNSTONE, I.M. (1994b). Minimax risk over  $\ell_p$ -balls for  $\ell_q$ -error. *Probab. Theory Related Fields.* 99, 277–303.
- [15] JOHNSTONE, I. and SILVERMAN, B. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* 32, 1594–1649.
- [16] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* 17, 1001–1008.
- [17] NICKL, R. and SZABÓ, B. T. (2014). A sharp adaptive confidence ball for self-similar functions arXiv:1406.3994.
- [18] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* 28, 298–335.
- [19] ROBINS, J. and VAN DER VAART, A.W. (2006). Adaptive nonparametric confidence sets *Ann. Statist.* 34, 229–253.
- [20] SZABÓ, B. T., VAN DER VAART, A.W. and VAN ZANTEN, J.H. (2014). Honest Bayesian confidence sets for the  $\ell_2$ -norm. arXiv:1311.7474.
- [21] SZABÓ, B. T., VAN DER VAART, A.W. and VAN ZANTEN, J.H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* 43, 1391–1428.
- [22] VAN DER PAS, S.L., KLEIJN, B.J.K. and VAN DER VAART, A.W. (2014). The horse-shoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Stat.* 8, 2585–2618.